1-1-1995

Performance of the Mantel-Haenszel, Simultaneous Item Bias and Logistic Regression procedures for detecting differential item functioning.

Pankaja Narayanan
University of Massachusetts Amherst

PERFORMANCE OF THE MANTEL-HAENSZEL, SIMULTANEOUS ITEM BIAS AND LOGISTIC REGRESSION PROCEDURES FOR DETECTING DIFFERENTIAL ITEM FUNCTIONING

A Dissertation

by

PANKAJA NARAYANAN

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF EDUCATION

May 1995

School of Education

PERFORMANCE OF THE MANTEL-HAENSZEL, SIMULTANEOUS ITEM BIAS AND LOGISTIC REGRESSION PROCEDURES FOR DETECTING DIFFERENTIAL ITEM FUNCTIONING

A Dissertation

by

PANKAJA NARAYANAN

Approved as to style and content by:

H. Swaminathan, Chair

Ronald K. Hambleton, Member

Gene A. Fisher, Member

Bailey W. Jackson, Dean
School of Education

DEDICATION

This work is dedicated to my family:

Ram Mohan, Sunder, Vijayalakshmi, Gitanjali and Jayashri

ACKNOWLEDGMENTS

I would like to express my sincere appreciation and thanks to many individuals for their assistance, encouragement, and support to accomplish this project.

First, I would like to convey my heartfelt appreciation and thanks to the members of my doctoral dissertation committee. I would like to extend my special thanks to Professor H. Swaminathan for being the chief motivating force behind this project. His guidance, encouragement, patience, and advice have been greatly instrumental in completing this study. I would like to give special recognition to Professor R. K. Hambleton for his active role, guidance, and support. His thoughtful comments have greatly enhanced the quality of this dissertation. Special thanks are also due to Professor Gene A. Fisher for his consistent support, encouragement, help, and invaluable contributions throughout the study period.

I also wish to thank Professor William F. Stout and his graduate students at the University of Illinois, Urbana-Champaign, for their assistance and guidance during the project.

I would like to extend my heartfelt thanks to Peggy Louraine for her efficient help in the enormous task of typing and editing this manuscript.

Thanks are also due to all my friends and fellow graduate students in Research and Evaluation program for their friendship, kindness, encouragement, and support. Finally, I would like to convey my appreciation to my family members for their unending support and understanding during my study.

ABSTRACT

PERFORMANCE OF THE MANTEL-HAENSZEL, SIMULTANEOUS ITEM BIAS AND LOGISTIC REGRESSION PROCEDURES FOR DETECTING DIFFERENTIAL ITEM FUNCTIONING

MAY 1995

PANKAJA NARAYANAN, B.A., UNIVERSITY OF MADRAS

M.A., UNIVERSITY OF MADRAS

M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

Ed.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor H. Swaminathan

The performance of three popular procedures for detecting differential item functioning (DIF), the Mantel-Haenszel (MH), the Simultaneous Item Bias (SIB), and the Logistic Regression (LR) procedures were investigated and compared in three different studies.

The first study compares the MH and the SIB procedures with respect to their Type I error rates and power to detect uniform DIF. Data for the study were simulated to reflect a variety of conditions. The results revealed that both the MH and the SIB procedures were equally powerful in detecting uniform DIF under most of the studied conditions. The SIB procedure showed higher detection rates than the MH procedure as the ability distribution differences increased.

The second study investigated the distributions of the SIB and two variations (with and without the continuity correction in the MH statistic), to determine whether or not their distributional assumptions held. The results showed that the SIB statistic generally had the expected distributions when the sample size of the reference and the focal groups exceeded 200. The distributions assumptions of the MH statistic without the continuity correction were more readily met than

those of the MH statistic with the continuity correction for all the studied conditions.

The third study investigated the MH, the SIB, and the LR procedures with respect to their Type I error rates and power to detect non-uniform DIF. Data for the study were simulated under a variety of conditions. The results revealed that both the SIB and LR procedures were equally powerful in detecting non-uniform DIF under most conditions. The MH procedure was not very effective in identifying non-uniform DIF items that showed disordinal interactions.

The investigation of the Type I error rates in all the three studies showed that they were within the expected limits for the MH procedure, higher than expected for the SIB and LR procedures with the SIB results showing an overall increase of about 1% over the LR results. With respect to power, the results show that the MH statistic was very effective in detecting only uniform DIF; both the SIB and LR procedures were very effective in detecting uniform as well as non-uniform DIF.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

C H A P T E R I

INTRODUCTION

Standardized achievement and ability tests are designed, administered and evaluated to serve many purposes. In educational settings, test results are used as criteria for making important decisions such as success or placement in schools and colleges, and diagnosis. In professional settings, they are often used as criteria for selecting candidates for certification, licensure decisions, and competency. To make decisions fair to all test takers based on test results, test scores should be capable of reflecting the true abilities of all examinees with a reasonable degree of accuracy. Additionally, no group of examinees should be at an advantage as compared to another group due to factors extraneous to the purpose of testing.

As the use of tests in making important decisions has increased, the possibility of bias in testing has been a major focus of test developers, administrators and researchers. The issue of bias arises from observed differences between the performances of different groups defined by ethnic background, gender and culture. Test developers have become increasingly concerned with litigation issues in courts of law over the use of tests in professional and educational settings. This has resulted in continued awareness on the part of the test developers in the test development process and the use of test scores.

Over the past two decades, a number of studies focused on both statistical and judgmental methods for detecting potentially biased items between two subgroups in standardized tests (see for example,

Berk, 1982; Hambleton & Rogers, 1989; Holland & Thayer, 1988; Hills, 1989; Rudner, Getson, & Knight, 1980a; Swaminathan & Rogers, 1990). Statistical techniques for investigating DIF mainly focus on three different kinds of approaches to investigate the best approximation techniques to assess test validity. The focus of these three approaches are studies conducted to investigate predictive validity, content validity and construct validity.

Predictive validity studies are undertaken to achieve fair selection opportunities in employment and college admissions. A number of methods to investigate predictive validity are presented in Petersen and Novick (1976). Predictive validity studies pertain to the functioning of the test as a whole and are therefore undertaken after the tests are constructed. On the other hand, content validity or construct validity studies are concerned with the internal structure of tests. All item bias detection techniques are statistical techniques focused to determine how accurately test scores represent the construct measured by the test.

Content validity studies pertain to the adequacy of the test items as a sample from a well-specified content domain and are associated with judgmental review procedures. In judgmental review methods, minority experts and judges identify potentially biased items. Often expert judges, although sensitive to specific instances of the cultural bias and stereotyping in tests, do not identify the same items which statistical techniques identify. Therefore, most researchers have not been very successful in their attempts to build judgmental strategies based on items identified by statistical item

detection procedures. In practice, studies comparing judgmental and empirical approaches have shown little agreement (Plake, 1980).

Many researchers prefer to use the more neutral term differential item functioning (DIF) rather than the term item bias because DIF focuses on what is empirically determined by statistical procedures about the relative performance of the two groups on the test items. The term item bias implies judgements about the qualitative aspects of test items in addition to DIF. Since this study is concerned with investigating techniques for determining empirically whether or not items function differentially for two groups, the term DIF will be used throughout this study.

Previous investigations on DIF have compared a number of statistical procedures for detecting DIF in efforts to identify and understand the best methods for detecting DIF (Hambleton & Rogers, 1989; Scheuneman & Bleistein, 1989; Rogers, 1989; Swaminathan & Rogers, 1990; Wright, 1986). These procedures differ from one another in many aspects: parametric or non-parametric, statistical and computational complexity, theoretical underpinnings, index vs. tests of significance based, and sensitivity. None of the methods have been found to be appropriate in all situations.

A researcher interested in selecting a DIF detection procedure is confronted with many methods with certainly no clear guidelines for choosing the best one for a given situation. Therefore, more research is needed in the measurement field to understand this important area.

Statement of the Problem

A review of literature on DIF reveals that approaches for detecting DIF can be divided into three broad areas. They are: methods based on classical test theory (CTT), methods based on item response theory (IRT), and methods based on chi-square techniques (Hills, 1989; Rogers, 1989; Scheuneman & Bleistein, 1989). Although a variety of procedures are available for DIF detection, these approaches appear to have limited scope for reaching general conclusions accounting for different mechanisms underlying the occurrence of DIF. Statistical studies conducted on DIF so far have typically not been able to produce results which could be generalized into offering guidelines for test developers and practitioners. However, for practitioners wanting to choose the DIF detection procedures appropriate to their datasets, it would be useful to know the strengths and weaknesses of different procedures. Therefore, empirical research comparing several statistical methods is necessary because issues such as sample size and other factors, computer execution time, and cost-effectiveness can clearly influence the results. If all DIF detection approaches tend to identify the same items as DIF, the most practical solution would be to use the simplest and the least expensive approach. However, if the approaches identify different items as DIF, it becomes necessary to use those methods which are most valid and stable. Therefore, with a plethora of DIF detection procedures currently available for detecting DIF, empirical research to compare the various methods is necessary to determine the conditions under which each procedure is optimal for detecting DIF.

Differential item functioning is said to exist if examinees of the same ability but from different subgroups have differing probabilities of answering an item correctly. There are two types of DIF that can occur in educational data. They are uniform and non-uniform DIF. Uniform DIF occurs when there is no interaction between group membership and the ability variable. Non-uniform DIF occurs when there is an interaction between group membership and the ability variable. In general, although uniform DIF occurs more often than non-uniform DIF, identification of non-uniform DIF items has been reported with real data (Hambleton & Rogers, 1989; Mellenbergh, 1982). Within the IRT framework, an item shows DIF if the item characteristics curves (ICCs) for the two groups are not the same. Therefore, the investigation of DIF in terms of IRT is a matter of comparing the ICCs for the two groups. Also, a clear distinction needs to be made while investigating uniform and non-uniform DIF. In the context of IRT, uniform and non-uniform DIF are represented by parallel and non-parallel ICCS, respectively.

For many years, IRT-based DIF detection procedures have been very popular due to their theoretical soundness. However, the major drawback of these procedures is that they require large sample sizes to produce independent stable estimates for both the reference and the focal groups. Moreover, IRT-based methods require complex computer programs and considerable expertise in running the programs. In addition, IRT indices such as the area between the ICCs for the two groups have no associated tests of significance. Because of this problem, in recent years, researchers have been involved in developing

methods, that are theoretically sound, computationally non-intensive, and cost-effective.

Currently, some of the popular non-parametric methods for detecting DIF are (1) the Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988), (2) the Standardization (STD) procedure (Dorans & Kulick, 1986) and (3) the Simultaneous Item Bias (SIB) procedure (Shealy & Stout, 1993). Swaminathan and Rogers (1990) presented the Logistic Regression (LR) procedure (1990) and demonstrated that despite its parametric nature, it is theoretically sound, effective, and easy to implement in practice. The focus of this study was an empirical comparison of the MH, SIB and LR procedures for detecting uniform and non-uniform DIF. The three procedures are described briefly below.

The MH procedure is one of the most popular and widely used procedure in DIF studies. Previous research comparing the MH procedure with other DIF detection procedures show that it is very effective in detecting uniform DIF (Hambleton & Rogers, 1989; Rogers & Swaminathan, 1993; Shealy & Stout, 1993; Swaminathan & Rogers, 1990). In this procedure, most typically, the raw score is used as the conditioning variable to form groups of examinees of comparable ability. For two groups matched on k+1 score categories, where k is the number of items in the test, the MH procedure compares the odds of success for the two groups of interest. The MH statistic provides both a test of statistical significance and a measure of DIF effect size (Holland & Thayer, 1988). It is very effective with small sample sizes and is also inexpensive to use in practice. However, one of the main limitations of this procedure is its incapability to detect non-

uniform DIF in some types of items (i.e., items of medium difficulty). This is because, the MH statistic being a signed statistic, is sensitive to the direction of DIF. When the direction of DIF changes in the middle of the ability distribution, negative differences in one part of the score distribution will cancel out the positive differences in the other. Therefore, non-uniform DIF items of this form may not be detected by the MH procedure.

The SIB procedure was developed based on the multidimensional IRT model of test bias. It emphasizes the examination of DIF at the test level and provides a statistical test to detect DIF present in one or more items on a test simultaneously. The SIB procedure requires the identification of a "valid" subtest for matching examinees. Although derived using IRT, the SIB procedure uses sample means and variances of scores of valid and studied subtests to obtain the test statistic. Like the MH statistic, the SIB statistic provides both a test of statistical significance and a measure of DIF effect size. It is computationally non-intensive, and effective with small sample sizes. The SIB statistic is designed to detect unidirectional and non-unidirectional DIF.

The LR procedure presented by Swaminathan and Rogers (1990) is a model-based procedure which allows for testing the hypothesis of no interaction between the ability variable and group variable. The major advantage of the LR procedure is that it can be expanded to condition on more than one test or subtest score.

From the brief descriptions of the MH, SIB and the LR procedures stated above, it is clear that both the MH and the SIB procedures share a common framework. Both procedures are non-parametric,

theoretically sound, and computationally simple. Studies comparing the MH and the SIB procedures show that both procedures are equally effective in detecting uniform DIF (Ackerman, 1992; Narayanan & Swaminathan, 1993; Roussos & Stout, 1993; Shealy & Stout, 1993).

Recently, a modification of the SIB procedure to detect non-uniform was introduced by Li and Stout (1993). The new SIB procedure, henceforth referred to as CRO-SIB, has also the potential for conditioning on more than one test or subtest scores. Because of its newness, the new SIB procedure for detecting non-uniform DIF has not been extensively studied. It is therefore appropriate at this stage to compare it with the LR procedure which at present, is known to be superior to other DIF detection procedures in detecting non-uniform DIF.

Previous research comparing the MH and the LR procedures show that both procedures are equally effective in detecting uniform DIF and that the LR procedure is more effective than the MH procedure in detecting non-uniform DIF. In fact, according to Swaminathan and Rogers (1990), the MH procedure can be conceptualized as being based on the LR model where the ability variable is treated as discrete and no interaction between the ability variable and group membership is allowed. The LR procedure would therefore be expected to improve on the MH procedure for detecting non-uniform DIF. Although a number of issues concerning DIF have already been resolved by previous research, a more thorough investigation of the three procedures under a variety of conditions was the main purpose of this investigation.

Purpose of the Study

The main purpose of this investigation was the comparison of three DIF detection procedures, the MH, SIB and LR, in order to determine their relative efficacy to detect DIF in test items. Three separate studies were conducted using simulated data. The first study was focused on an empirical comparison of the MH and the SIB procedures for their capability to detect uniform DIF. The second study investigated the asymptotic distributional properties of the MH and the SIB statistics. The third study was an empirical comparison of the MH, SIB and the LR procedures in terms of their capability to detect non-uniform DIF. A brief description of each of the three studies is given below:

1. The purpose of the first study was to compare the MH and the SIB procedures with respect to their Type I error rates and power to detect uniform DIF. The study investigated the conditions under which each procedure was optimal for detecting uniform DIF. Data for the study were simulated to reflect a variety of conditions: the factors manipulated were sample size, ability distribution differences between the reference and the focal groups, proportion of items showing DIF, DIF effect size, and the type of item.
2. The second study investigated the distributional assumptions of the MH and the SIB statistics to determine the conditions under which their asymptotic distributions were obtained. Previous research investigating the distributional assumptions of the MH statistic has shown that they were not satisfied for many

conditions (Rogers, 1989; Narayanan & Swaminathan, 1993). These results raise questions about the practice of using the MH statistic (with the continuity correction) as a test statistic for detecting DIF. Using simulated data, the study therefore investigated the distributional assumptions of SIB statistics and two variations of the MH statistic (with and without the continuity correction). The power and Type I error rates of the SIB and the MH statistic (with and without the continuity correction) were investigated to determine the effect of the continuity correction on the statistic.

3. In the third study, the power and Type I error rates of three DIF detection procedures, the MH, SIB and the LR procedures were investigated to determine their capability to detect non-uniform DIF. Previous research has indicated that the MH procedure was not capable of detecting non-uniform in certain types of item whereas the LR procedure was very effective in detecting such items. The SIB procedure developed by Li and Stout (1993) for detecting non-uniform DIF is relatively new and therefore comparison with an already effective procedure as LR would be timely. The study was conducted with simulated data to reflect a variety of conditions. The factors manipulated were sample size, ability distribution differences, proportion of items containing DIF, DIF effect size and type of item.

There are six chapters included in this dissertation. Chapter I provides a brief introduction and the purpose of the study. In Chapter II, a review of the literature relating to the detection of

DIF is presented. The research objectives, methodology and the results for each of the three studies are presented in Chapters III through V. In Chapter VI, a summary of the study and the conclusions drawn from the results of the study are given.

C H A P T E R II

LITERATURE REVIEW

For many years, differential item functioning (DIF) between different ethnic, cultural and gender groups has been one of the major areas of research in the field of educational measurement. A review of literature on DIF indicates that research in this area over the years has yielded a variety of statistical procedures for detecting DIF (see for example, Berk, 1982; Rogers, 1989). In this chapter, DIF is briefly defined, followed by a summary and description of some of the prominent statistical methods for detecting DIF.

Definition of Differential Item Functioning

Differential item functioning (DIF) can occur when there are observed differences in performance between two different subgroups of interest. The current definition of DIF states that an item is differentially functioning if examinees from different groups but of the same ability have different probabilities of answering an item correctly.

Most definitions of DIF can be classified under one or other of two general headings. They are: (1) definitions that are unconditional on ability, and (2) definitions that are conditional on ability.

In unconditional methods, an item is defined as functioning differently for two groups if the item is relatively easier for one group than the other, i.e., if there is item by group interaction. Unconditional methods depend on the score distributions of the two

groups of interest. Therefore, in these methods, DIF is dependent on the other test items.

In conditional methods, an item is defined to be functioning differently for two groups if the probability of a correct response is not the same for the two groups at a given ability level. These methods, therefore, are not dependent on score distributions of the two groups of interest. Therefore, in these methods, DIF is independent of the other test items (Mellenbergh, 1982).

When differences in performance exist between two groups, a distinction needs to be made between DIF and impact. Impact is said to be present when the performance differences that can occur are only due to consistent and stable differences in examinee ability distributions across groups. Impact therefore reflects the differences in the overall ability distributions across groups. In contrast to impact, DIF is said to exist if, after matching the two groups with respect to the ability that the test is assumed to measure, the item shows differential functioning for the two groups.

The definition of DIF in terms of the probability of a correct response can be restated in terms of an item response theory framework. From the IRT perspective, a test item is differentially functioning if the item characteristic curves (ICCs) across different subgroups are identical (Hambleton & Swaminathan, 1985). Therefore, the investigation of DIF within the IRT framework is a matter of comparing the ICCs for the two groups of interest.

There are two types of DIF, that can occur in educational data, uniform and non-uniform DIF. Uniform DIF occurs when there is no interaction between ability level and group membership. That is, the

probability of answering an item correctly is greater for one group than the other uniformly over all ability levels. Non-uniform DIF can occur when there is interaction between ability level and group membership. That is, the difference in the probabilities of answering an item correctly is not the same at all ability levels. In terms of IRT, uniform and non-uniform DIF are represented by parallel and non-parallel ICCs respectively.

Statistical Procedures for Detecting DIF

Statistical DIF detection methods are designed to detect DIF either by unconditional or conditional methods. There are three major approaches for detecting DIF among test items. These approaches can be classified as follows:

1. Classical Test Theory Methods - In classical test theory methods, basically the observed test score is used for comparison between two groups rather than the "true" scores under the assumption that the observed test score is a valid and reliable measure of examine ability. These methods mainly compare the item difficulties (p-values) and to a lesser extent compare the item discriminations (r-values) for the two groups. However, these methods are sample dependent.
2. Item Response Theory Methods - Item response theory methods for detecting DIF which involve comparison of the ICCs for the two groups to investigate if they can provide statistical evidence of the differences between the ICCs for the two groups of interest.

3. Chi-Square Methods - Chi-square methods for detecting DIF are based on the construction of two-way contingency tables (group by item response). These procedures use the chi-square indices to test the null hypothesis of no DIF between the two groups.

4. Log-linear Methods and the Logistic Regression Procedure - Log-linear methods are chi-square methods for detecting DIF based on the construction of three way contingency tables (score category by group by item response). That is, examinees are sorted into subgroups of the same total score, and two-way contingency tables (group by item response) are constructed. These procedures can be thought of as an extension of regression analysis, where both the dependent and the independent variables are discrete. The logistic regression procedure is a DIF detection procedure based on the logistic regression model.

Classical Test Theory Approaches

The three prominent methods for detecting DIF in this approach are (1) the ANOVA method, (2) the transformed item difficulty method (Angoff, 1982), and (3) the standardization method (Dorans & Kulick, 1986). The ANOVA and the transformed difficulty methods are based on the unconditional methods for detecting DIF whereas, the standardization method is based on the conditional methods for detecting DIF. In the following sections, the ANOVA method and the transformed item difficulty methods will be described. The standardization method will be described in a later section.

The Analysis of Variance Method. The analysis of variance method (ANOVA) for detecting DIF involves performing the statistical

procedure, viz., analysis of variance with item and group membership as the independent variable and the item score as the dependent variable. If the results show that item by group interaction is significant, it can be then concluded that there is statistical evidence of the presence of DIF in certain items in the test. This is an indication that some of the test items are relatively more difficult for one group than the other. Before conducting analysis of variance, an arcsin transformation of the item difficulties or the p-values for the two groups (Cardall & Coffman, 1964), needs to be effected to satisfy the assumptions of homogeneity of variance required by the ANOVA model.

Research studies show that the ANOVA method for detecting DIF has many limitations. Lord (1980) pointed out that results obtained from comparisons of p-values for the two groups can be misleading. When there are real differences in group performance, it is likely that highly discriminating items will be able to distinguish better between groups and tend to show larger difference in percentage correct scores for an item and falsely label the item as DIF for the two groups. Also in comparisons between groups of differing abilities, differences in difficulties between the two groups can also produce significant item by group interactions and can result in item being falsely labeled as DIF (Rudner et al., 1980a; Camilli & Shepard, 1987). Therefore, the ANOVA method which provides a test of significance to determine the differences in p-values may not be appropriate in detecting DIF. Camilli and Shepard (1987) also showed that ANOVA may fail to detect items that functioning differently for members of two groups when true differences exist between the two

groups. In their simulation study, they showed that when there are real group differences, DIF may not be detected because the contributions to the between-group variance than to the variance due to group by item interaction and as a result DIF may not be detected when there are real occurrences of DIF in the test items.

Several studies conducted based on ANOVA method indicate that this procedure cannot be effective as a method for detecting DIF (Cardall & Coffman, 1964; Cleary & Hilton, 1968; Camilli & Shepard, 1987). Camilli and Shepard suggest that ANOVA should no longer be recommended as a DIF procedure, even during the preliminary screening of test items.

Transformed Item Difficulty Method. Like the ANOVA method, the transformed item difficulty (TID) procedures are based on the assumption that differential item functioning is the effect of item-group interaction. In interaction procedures, an item is considered biased if compared to other items on the test, it is relatively more difficult for one group than the other.

The most popular of the TID procedures is the delta plot method (Angoff, 1982; Angoff & Ford, 1973). This method involves computing the item difficulties or the p-values separately for each item for the two different groups of interest. Lord (1980) has shown that the relationship between the item difficulties or the p-values is non-linear. Therefore, before applying the delta-plot method, the p-values are first normalized for each group by computing the normal deviate z corresponding to the (1-p)th percentile of the normal distribution. To eliminate the negative z-values, they are then converted to delta values with a mean of 13 and standard deviation

equal to 4. There will be a pair of delta values for each item for the two groups. The pairs of delta values for the two groups are then plotted on a graph. When the two groups are of equal abilities, the delta points when plotted, will form an ellipse extending from the lower left to the upper right along the 45^{o} line passing through the origin representing the line of equal difficulty. When the two groups differ in abilities, the delta points will be tilted vertically or horizontally from the 45^{o} line.

From the scatterplot of the delta points, a straight line of best fit (which minimizes perpendicular deviations) can be fitted which represents the major axis of the ellipse formed by the delta points. The deviations of a given point from the major axis line is taken as a measure of bias for that item. Points that fall further away from the major axis line represent items that contribute item by group interaction and are items that are relatively more difficult for one group than the other.

Angoff and Ford (1973) provide formulas for determining the equation to the major axis line of the ellipse and for computing the distance of each point from the line. The equation to the major axis of the ellipse is given by $Y = AX + B$ where

$$A = \frac{(S_y^2 - S_x^2) - \sqrt{[(S_y^2 - S_x^2)^2 + 4r_{xy}^2 S_x^2 S_y^2]}}{2r_{xy} S_x S_y} \qquad (1)$$

and

$$B = M_y - AM_x \qquad (2)$$

where x and y respectively are the delta values for the two groups of interest, where M_x and S_x refer to the mean and the standard deviation respectively of the deltas for the group plotted on the x-axis, and r_{xy}

refers to the correlation between the two groups. The formula for the perpendicular distance D_i of each point i in the plot to the line is given by

$$D_i = \frac{Ax_i - Y_i + B}{A^2 + 1} \qquad (3)$$

Echternacht (1974) describes a variation of the delta-plot method to assess differential item functioning. In this method, the item difficulties or the p-values are transformed to delta values. The differences between the corresponding delta values for the two groups are computed and the distributions of the differences are tested for normality. Echternacht states that within sampling limits, the differences between paired delta values will be constant across all the items when there is no differential item functioning.

Another variation of the delta-plot method was proposed by Coffman (1961, 1963). In this method, item difficulties for the two groups are computed and paired values of $2\arcsin\sqrt{p}$ are plotted, where p is the proportion of the two groups answering the item correctly. The advantage is that the transformed item difficulty values will have the same variance error (1/N), where N is the number of cases in the sample. The disadvantage is that, like the original p-values, these values are bounded and therefore, yield a curvilinear plot when the two groups under study have different means (Angoff, 1982).

Sinnott (1980) proposed a iterative modified method to determine the major axis line to decrease the effect of items functioning differently for the two groups while computing the major axis line. In this method, the major axis line based on all the items is

determined. Items identified as differentially functioning based on their distances from the line are removed. A new line is then determined based on the remaining items. The removed items are then readmitted and the process is continued until the same items are identified in two consecutive "purifications."

Rudner et al. (1980a) proposed a modification of the delta-plot method to directly compare the p-values. In this analysis, item p-values are computed for each group separately and transformed to within-group z-values using the item mean and standard deviations for that group. The plot of paired z-values which yields a major axis line with the 45° line is used as the reference line against which item-point discrepancies are measured. The perpendicular distances of the paired z-values from the 45° line are used to indicate the magnitude of bias.

The advantages of the delta-plot method are that they are simple, inexpensive and do not require large numbers of examinees although a large number of items are required to obtain a well-defined major-axis line. There are some limitations to the delta-plot method. When the two groups differ in their mean ability, larger item difficulty differences will be obtained by highly discriminating items whereas smaller item difficulty differences will be obtained by low discriminating items even when the items are not differentially functioning. Therefore, this method may identify as differentially functioning items that are simply highly discriminating (in relation to other items) and not differentially functioning. Angoff (1982) has suggested using groups matched on ability before plotting the p-values.

Overall, the transformed item difficulty procedures can provide results closely approximating those of IRT methods by suitable adjustments for ability differences in the two groups and in item-test correlations. They have been highly recommended for use at least for small samples where other methods cannot be used.

Item Response Theory Approaches

Currently, the most popular methods for detecting DIF are the statistical methods based on item response theory (IRT). IRT allows examine performance on a set of test items to be expressed as a function of one or more characteristics referred to as traits or abilities. In other words, IRT models specify a mathematical function expressing a relationship between the observable examine test performance and the unobservable traits or abilities responsible for examine test performance. For each item in the test, a monotonically increasing curve called the item characteristic curve (ICC) represents the conditional probability function (logistic or normal ogive) of a correct response to the item for a given level of ability. The assumption that there is one dominant ability that explains examine performance is fundamental to item response theory.

There are several IRT models currently being used in the design and analysis of educational and psychological data. The principal difference among the models is the manner in which the mathematical form of the ICCs are specified and the number of parameters required to describe the items. Some of the IRT models include (1) unidimensional and multidimensional models, (2) linear and non-linear models, and (3) dichotomous and polychotomous models.

The most general form of the three popular logistic models is the three-parameter logistic model (Birnbaum, 1968). The mathematical form of the three-parameter logistic model is given by

$$P_i(\theta) = c_i + (1 - c_i) \frac{\exp(Da_i\ (\theta - b_i))}{1 + \exp(Da_i\ (\theta - b_i))} \qquad (4)$$

where

$P_i(\theta)$ = the probability that an examine with ability θ will answer an item correctly

b_i = the item difficulty parameter

a_i = the item discrimination parameter

c_i = the pseudo-chance level parameter

D = 1.7 (a scaling factor)

The difficulty parameter b_i corresponds to the point on the ability scale at which the probability of a correct response is equal to (1+c)/2. It is located at the point on the ability scale where the slope of the ICC is maximum. High values of the difficulty parameter represent items that are relatively more difficult than others. The discrimination parameter a_i corresponds to the slope of the point where the difficulty parameter is located on the ability scale. High values of the discrimination parameter a_i have steep slopes and represent items that are more effective in distinguishing between high and low ability examinees. The pseudo-chance level parameter c_i represents the lower asymptote of the ICC and corresponds to the probability of correct response for examinees whose ability levels are very low. Items for which much guessing is involved have high values

of pseudo-guessing level parameter. The three-parameter model is the most popular of the IRT models.

The two-parameter IRT model proposed by Birnbaum (1968) takes the form

$$P_i(\theta) = c_i + (1 - c_i) \frac{\exp(Da_i\ (\theta - b_i))}{1 + \exp(Da_i\ (\theta - b_i))} \qquad (5)$$

The two-parameter logistic model is a special case of the three-parameter model in which all the c-values are zero. In such circumstances, guessing is not likely and consequently, very low ability examinees have little or almost no chance of responding positively to difficult items. The one-parameter logistic model has the mathematical form given by

$$P_i(\theta) = \frac{\exp(Da(\theta - b_i))}{1 + \exp(Da(\theta - b_i))} \qquad (6)$$

The one-parameter logistic model is the simplest of the IRT models, least expensive, easy to use and involves fewer item parameters. The one-parameter is obtained from the two parameter model by setting all the a-parameter values to a constant. One of the main limitations of the model is that it is more difficult to fit the one-parameter model to the data and therefore, the use of the model is restricted.

There are three major advantages of the IRT model over CTT models (Hambleton & Swaminathan, 1985). They are: (1) the descriptors of the test items (item parameters) are independent of a particular sample of examinees chosen for calibrating the item, (2) the estimate

of an examinee's ability is independent of a particular choice of test items, and (3) a standard error indicating the precision with which an examine's ability is estimated is obtained.

The IRT method for detecting DIF involves comparison of ICCs for the two groups. Therefore, in IRT, an item is not functioning differently for two groups, if the estimated ICCs for the two groups should be identical within the limits of sampling errors. If the ICCs for the two groups, after being placed on the same scale, differ for reasons other than sampling errors, then, the item is said to be functioning differently for the two groups. Therefore, investigation of DIF in IRT involves measuring the extent to which the ICCs vary across the two groups. There are two steps: (1) the item parameters are estimated for each group and expressed on the same scale, and (2) a index of DIF is computed for each item.

IRT Methods for Detecting DIF

A review of the DIF methods based on IRT is given in Shepard, Camilli & Averill (1981) and Ironson (1982). The three major procedures for detecting DIF based on IRT (Hambleton & Swaminathan, 1985) are:

1. Comparisons of item characteristics curves for the two groups,
2. Comparison of the vectors of the item parameters, and
3. Comparison of the fit of the IRT models to the data.

Comparison of Item Characteristic Curves. Rudner (1977) proposed a method to determine the area between the ICCs of the two groups of interest. After choosing an appropriate IRT model, the item

and ability parameters are estimated for the two groups. The item parameters are then placed on the same scale by standardizing on the b_i's. Standardizing the item parameters on θ's requires a scale equating procedure. The ability scale, usually defined from -3 to +3 is then divided into intervals of width $\Delta\theta$ (= .005). After determining the value of θ_k at the midpoint of the ability interval k, the height of the two item characteristics curves $P_{i1}(\theta_k)$ and $P_{i2}(\theta_k)$ at θ_k are computed. The area between the estimated item characteristic curves for the two groups is given by

$$A = \sum_{\theta_i=-3}^{\theta_i=+3} \left| P_{i1}(\theta_k) - P_{i2}(\theta_k) \right| \Delta\theta \qquad (7)$$

A small value obtained for A for an item indicates that the area between the two curves is small and therefore differential functioning for the item is small (Hambleton & Swaminathan, 1985).

The above formula uses the absolute value of the difference between the curves at each ability level. The area A calculated between the two ICCs without taking into account the direction in which the item is differentially functioning, is an unsigned measure of DIF. If the difference between the two ICCs is taken as positive if the reference curve is above the focal curve, and negative if the focal curve is above the reference curve at each ability level, then the resulting area will be a signed measure of DIF (Ironson, 1982). The signed measure of DIF is given by

$$B \text{ (signed)} = \sum_{\theta_i=-4}^{\theta_i=+4} [P_{i1}(\theta_k) - Pi_{i2}(\theta_k)] \Delta\theta \qquad (8)$$

The choice of the interval over which DIF is assessed is important in IRT methods of detecting DIF (Hambleton & Rogers, 1989) and should be chosen to be appropriate with the study purpose. Hambleton and Rogers (1989) computed the area statistics for different ability intervals to assess the discrepancy in the score distributions of two groups. They concluded that, when the interval over which the area statistic was focused on the region of the scale consisting of the focal group members, a fewer number of non-uniformly differentially functioning items were identified.

Raju (1988) provides formulas for the exact signed and unsigned areas between the two item characteristics curves. He points out that when the lower asymptotes (c-parameters) are equal, then the areas between the two ICCs will be finite and can be estimated by integrating between two finite points on the θ scale. For unequal lower asymptotes, the area between the two ICCs will be infinite. It follows that the finite interval procedures for estimating the area between the two ICCs with unequal lower asymptotes yield misleading results.

Raju (1990) determined the standard error of the area statistic when the c-parameters are the same for the two groups. The area statistic between the ICCs for the two groups divided by the standard error has an approximate normal distribution.

The area method does not provide an associated test of significance to determine the area statistic value that can be used to flag an item as differentially functioning for the two groups. In the absence of any statistical test of significance, an alternate procedure to investigate for DIF is to establish a "cut-off" to

determine the value of the area statistic. For this purpose, a baseline for comparison is established by dividing the reference group into two randomly equivalent samples. After estimating the ICCs separately for the two groups, the area between them is computed. Since the area between the two randomly equivalent samples is expected to be equal to zero, differences in the area statistic values would indicate that sampling errors may be present. A significant value of the area statistic as an indication of DIF for the two groups would be the value larger than the "cut-off" value.

Rogers and Hambleton (1989) used a simulated set of data to determine the "cut-off" value. In this method, the data sets are simulated for the reference and the focal groups using the item parameters estimated from the real data. Item and ability parameters are separately estimated for each set of simulated data for the two groups. Since there is DIF present, area values that are not equal to zero may be taken to be due to sampling errors. As mentioned before, the largest area value from this comparison may be used as a cut-off value to compare if DIF is present for the two groups in the real data set.

Linn, Levine, Hastings and Wardrop (1981) proposed a method which involves weighting the b-parameters so that their weighted mean and variance are equal for the two groups. Their DIF statistic is given by

$$C = \sum_{\theta_i=-3}^{\theta_i=+3} \{[P_{i1}(\theta_k) - P_{i2}(\theta_k)]^2 \ \Delta\theta\}^{\frac{1}{2}} \qquad (9)$$

In this method, the ability scale between -3.0 to +3.0 is divided into 600 intervals of width equal to 0.01 each. At the midpoint of each interval, the difference between the ICCs for the two groups is squared, multiplied by the width (0.01), summed and the square root of the sum is taken as the index of DIF. Although this index is similar to the area method, it has a sampling distribution associated with it so that a significance test can be performed to determine whether the two curves differ by more than what it would be due to chance alone.

Comparison of Item Parameters. If two item characteristic curves differ across groups, then the estimates of the item parameters for the two groups would also differ. Equality of item parameters can be examined separately or simultaneously to establish that an item is functioning differently for the two groups (Hambleton & Swaminathan, 1985).

Lord (1980) proposed an asymptotic significance test for comparing the a and b parameters between two different groups. In this method, the two groups are combined and the item parameters are estimated standardizing on the b_i's. The c_i's are then fixed at the values obtained for the combined sample and the a_i's and the b_i's are reestimated separately for the two groups standardizing on the b_i's. After placing the a_i's and the b_i's on a common scale, they are compared to determine if differences exist in the ICCs for the two groups. A chi-square test with two degrees of freedom is provided for the purpose of comparing the two ICCs.

Hambleton and Swaminathan (1985) point out that since Lord's asymptotic distribution is not known, it is not possible to determine

the sample size to enable the asymptotic distribution to hold. Also, it is not known if the asymptotic distribution will hold when the item and ability parameters are simultaneously estimated.

McLaughlin and Drasgow (1987) in a computer simulation study examined the properties of Lord's chi-square method. They point out that when the item parameters of the IRT model estimated by the maximum likelihood method provides a reasonably good fit to the data, Lord's DIF statistic would be closely distributed as a chi-square distribution with two degrees of freedom. On the other hand, when person parameters are unknown and estimated simultaneously with item parameters, Lord's DIF statistic may not follow the chi-square distribution well enough to allow valid tests of DIF.

Wright, Mead and Draba (1976), suggested a Z-statistic to detect item DIF when item and ability parameters are separately estimated for the two groups after standardizing on the b_i's. The test statistic is calculated as follows:

$$Z = (b_{i1} - b_{i2}) / \sqrt{(SE_{i1}^2 + SE_{i2}^2)} \qquad (10)$$

where b_{i1} and b_{i2} are the estimated item difficulty for the ith item in groups 1 and 2, and SE_{i1} and SE_{i2} are the corresponding standard error of bi's in the two groups. The calculated Z-statistic can be compared with the tabulated standardized normal curve values and assessed if the item is differentially functioning for the two groups.

Comparison of Model Fit. Another procedure for detecting DIF is to compare the fit of the ICCs for the two groups of interest. When items are differentially functioning for the two groups, the assumption of unidimensionality is violated since examine performance

is affected by ability and group membership. Therefore, the models for the two groups will not fit each other.

Linn and Harnisch (1981) suggested a DIF procedure for comparing the fit of item response models in the two groups. The procedure is used when the sample size of the focal group is smaller than the sample size of the reference group. In this procedure, the samples of the reference and focal groups are combined together and the item parameters a, b and c for each item and for each examine are estimated. Following this estimation, the target group (i.e., the group against which DIF is suspected) is selected. From the item parameter and ability estimates, P_{ij}, the estimated probability that a person j would answer an item i correctly computed for each person in the target group. The quantity, P_{ig}, the proportion of examinees in subgroup g expected to get item i correct, and the quantity O_{ig}, the observed proportion correct on item i for subgroup g are then computed. The quantity $P_{i.}$, the proportion of examinees in the complete target group, and the quantity $O_{i.}$, the observed proportion correct on item i for the target group, are also computed. The difference $D_{i.} = O_{i.} - P_{i.}$ is an index of the degree to which members of the entire target group perform better or worse than expected on that item (Linn & Harnish, 1981).

Another index is a standardized difference score for each item for each subgroup. Here, the difference between the observed and expected performance for the group is expressed in standard deviation units under the assumption that DIF is zero. The formula used to calculate this is:

$$Z_{ig} = \frac{1}{N_g} \sum_{j \epsilon g} \frac{U_{ij} - P_{ij}}{\sqrt{[P_{ij}(1-P_{ij})]}} \qquad (11)$$

where

U_{ij} = 1 if person j answers item i correctly and 0 otherwise,

N_g = the number of persons in group g, and

P_{ij} = the estimated probability that person j would answer item i correctly, based on the fitted model from the combined groups.

Here, the difference between the observed and expected performance for the group is expressed in standard deviation units under the assumption that DIF is zero.

A modified form of the pseudo-IRT method was developed by Hoover and Kolen (1984). In this method, the item parameters are estimated after combining the two groups using the three-parameter model. Then the difference between the examinee's actual response and the expected probability of answering an item correctly (p) is computed separately for the two groups. This difference is then divided by $[p(1-p)]^2$, and these values are then averaged over the examinees in each group. The mean of each group is then squared and summed to provide an overall index of DIF.

Some of the advantages and disadvantages of the IRT-based models are given in Ironson (1982). Since IRT is a theoretically sound approach due to its property of item invariance, IRT-based models for detecting DIF are the most efficient. The invariance property of IRT models ensures that differences in ability do not create spurious instances of DIF. Also, item parameters are less confounded with

differences in item discrimination and guessing unlike classical p-values. The methods based on the three-parameter model are more efficient due to the inclusion of the discrimination and guessing parameters in the model.

The main disadvantage is that IRT-based methods are complex, expensive and require large sample sizes to estimate the parameters of the model. In addition, many indices such as the area between the two ICCs have no associated tests of significance.

Chi-Square Methods

Scheuneman (1979), Camilli (in Shepard et al., 1981), have proposed contingency table methods from which chi-square values can be computed and used as indices for DIF. In chi-square techniques, it is assumed that the test is reliable, valid and homogeneous so that the total test score may be used as an estimate of ability. Chi-square methods are therefore, considered as approximations to IRT methods since the observed test score is substituted for the ability. One of the advantages of these techniques is that the scores do not have to be normally distributed (Ironson, 1982). Different chi-square indices are obtained as a result of specifying different null and alternate hypotheses (Marascuilo & Slaughter, 1981).

Scheuneman's Chi-Square Index. Scheuneman (1979) proposed a method which involves constructing 2 x 2 contingency tables (group by item response) to compute the chi-square DIF index. For each test item, one such table is constructed to accommodate group by item response at each score level. In all, K 2 x 2 contingency tables are constructed where K is the distinct number of score levels formed for

the test. The 2 x 2 contingency table for the ith item and score level j is constructed in the form shown below:

	Score on studied item		
	1	0	Total
Group			
Reference	A_j	B_j	N_{Rj}
Focal	C_j	D_j	N_{Fj}
Total	M_{1j}	M_{0j}	T_j

In Scheuneman's method, the total score range is divided into five score or ability levels. The intervals must be specified to ensure that there are sufficient number of cases per cell in each score level. Within each score level, the probability of answering an item correctly should be the same for the two groups and should be < 1 in the lowest score interval and > 0 in the highest score interval. For each item, at each score level, a 2 x 2 contingency table is constructed. The expected frequencies of correct responses using the marginal totals are computed for each group. The observed frequencies minus the expected frequencies are computed and summed across all score levels. Scheuneman's chi-square goodness of fit is given by

$$\chi^2 = \sum_{j=1}^{5} \frac{((A_j - E(A_j))^2}{E(A_j)} + \frac{((C_j) - E(C_j))^2}{E(C_j)} \qquad (12)$$

where

$$E(A_j) = N_{Rj}M_{1j}/T_j$$

and

$$E(C_j) = N_{Fj}M_{1j}/T_j$$

For the Scheuneman's chi-square statistic, both signed and unsigned DIF indices can be computed. When uniform DIF is present, the signed

and unsigned indices will result in the same conclusions. When non-uniform DIF is present, then the signed index will not reflect DIF found by the unsigned chi-square index.

The chi-square index proposed by Scheuneman has an approximate chi-square distribution with N-1 (= 4) degrees of freedom where N (= 5) is the number of score groups. Several researchers have criticized Scheuneman's procedure on the grounds that the statistic has only an approximate chi-square distribution because the computations involved in computing the chi-square statistic uses only the correct responses (Baker, 1981; Camilli, 1979).

Camilli's Full Chi-Square Index. Camilli(1979) modified Scheuneman's method by computing the correct and incorrect responses by the two groups of interest in each score level for a 2 x 2 contingency table at each score level and then summed across all score levels. Camilli's chi-square statistic is given by

$$\chi^2 = \sum_{j=1}^{5} \frac{((A_j-E(A_j))^2}{E(C_j)} + \frac{((C_j)-E(C_j))^2}{E(D_j)} + \frac{(B_j)-E(B_j))^2}{E(C_j)} + \frac{((D_j-E(D_j))^2}{E(D_j)} \qquad (13)$$

Baker (1981) pointed out that the procedure is confounded by unequal sample sizes for the two groups. He also demonstrates that when the expected proportion of correct responses lies in the center of the two observed proportions, a large difference between two observed proportions at a given score level is not likely to result in a large contribution to the overall χ^2 statistic.

Crocker and Algina (1986) argued that with the chi-square techniques, artifacts of measurement errors may provide evidence of DIF. According to these authors, if comparisons are not made within specified score levels, then it amounts to comparing the item difficulty or proportion correct for the two groups. When true differences are present between the two groups in the same observed score level, controlling the observed scores in specified score levels may result in more frequent correct responses to an item in the group with higher true scores responding more frequently than the other group thereby contributing to measurement error.

Marascuilo and Slaughter (1981) point out that the null and alternate hypotheses specified by the two chi-square procedures are different. Scheuneman's chi-square method tests the null hypothesis that there are no group differences in proportion correct separately for each score group against the alternate hypothesis that there is a difference in each score group. On the other hand, Camilli's chi-square method test the null hypothesis that there are no group differences in proportion correct in any score group against the alternate hypothesis that there is a difference in at least one score group. Because of this Scheuneman's chi-square is likely to produce a large Type II error.

The chi-square techniques are advantageous because they are simple, inexpensive, require relatively small sample sizes and very limited computer time. Both chi-square procedures are associated with significance tests that allow dichotomous classification as differentially functioning and non-differentially functioning.

Marascuilo and Slaughter (1981) present two other chi-square methods in which different null and alternate hypotheses are specified. The first method tests the null hypothesis that there are no group differences in proportion correct at any score level against the alternate hypothesis of a constant group difference across score levels. This method would be sensitive to uniform bias only. The second method tests the null hypothesis that the group differences in proportion correct at all score levels is equal to a constant against the alternate hypothesis that the null hypothesis is false. This method would be more sensitive to non uniform bias than the first method. According to the Marascuilo and Slaughter, the latter method is essentially equivalent to a loglinear method.

The Mantel-Haenszel Procedure. The Mantel-Haenszel procedure (Holland & Thayer, 1988) compares the probabilities of a correct response in the two groups of interest for examinees of the same ability. In order to compare the probabilities of a correct response, item response data for the reference and the focal group members are arranged into a series of 2 x 2 contingency tables (group by item response) at each score level as described earlier.

The null DIF hypothesis of interest is that the odds of getting the item correct at a given score level j is the same for the reference and the focal group, at all K levels of the matching criterion. The null and alternate constant odds ratio hypothesis at score level j can be expressed as follows:

$$H_0: [\pi_{Rj}/(1-\pi_{Rj})] = [\pi_{Fj}/(1-\pi_{Fj})] \qquad j = 1, 2, \ldots, k \qquad (14)$$

$$H_a: [\pi_{Rj}/(1-\pi_{Fj})] = \alpha[\pi_{Rj}/(1-\pi_{Fj})] \qquad j = 1, 2, \ldots, k, \ \alpha \neq 1 \qquad (15)$$

where

π_{Rj} = the probability that a reference group examine with total score j will get the studied item correct

π_{Fj} = the probability that a focal group examine with total score j will get the studied item correct

The parameter α is called the common odds ratio. When the value of α is equal to one, the probability of a correct response is equal for both groups. A value of α greater than one indicates that the reference group members are more likely to answer the item correctly. Similarly, a value of α less than one indicates that the focal group members are more likely to answer the item correctly. An estimate of the common odds ratio α, known as MH-Alpha, also provides an estimate of DIF effect size. It can be expressed as

$$\text{MH-Alpha} = \frac{\Sigma A_j D_j / T_j}{\Sigma B_j C_j / T_j} \tag{16}$$

From the K 2 x 2 tables presented above, for a given item, the Mantel-Haenszel statistic with a continuity correction is computed as follows:

$$\text{MH-Chisquare} = \frac{(|A_j - \Sigma E(A_j)| - .5)^2}{\Sigma \text{Var}(A_j)} \tag{17}$$

where

A_j = the observed number of examinees in the reference group at score level j answering the item correctly,

$$E(A_j) = (N_{Rj} M_{1j}) / T_j \tag{18}$$

and

$$\text{Var}(A_j) = \frac{(N_{Rj} N_{Fj} M_{1j} M_{0j})}{(T_j)^2 (T_j - 1)} \tag{19}$$

The MH Delta statistic introduced by the Educational Testing Service (ETS) is a statistic which is obtained by a non-linear transformation of MH Alpha. A positive value of MH Delta can be interpreted as indicating that the item was easier for the focal group than for the reference group. MH Delta is given by

$$\text{MH Delta} = -(2.35)\ \ln\ (\text{MH Alpha}) \qquad (20)$$

The Simultaneous Item Bias Procedure. The Simultaneous Item Bias (SIB) procedure developed by Shealy & Stout (1991) emphasizes the examination of DIF at test level and provides a statistical test to detect DIF present in one or more items on a test simultaneously.

To test whether a set of items on the test is functioning differently, item response data for the reference and focal groups are formed into two subtests, a "suspect" subtest containing the items that are to be tested for DIF (they can be one or more items), and a "valid" subtest containing the items that measure the construct that the test is purported to measure. To calculate the SIB statistic, examine response data on the "valid" subtest scores are used to group the reference and focal groups into score levels so that, for n items in the test, the number of score levels on the "valid" subtest score will be equal to n+1. The reference and focal group members with the same valid subtest scores are then arranged to form statistic calculation cells such that each statistic calculation cell will correspond to a particular "valid" subtest score. Within each statistic calculation cell, the average proportion right on the "suspect" subtest is calculated for the reference and the focal groups.

Shealy and Stout's DIF index, β, is a parameter denoting the amount of DIF. For example, a β value of 0.1 indicates that the average difference in the probabilities of correct response of "studied" subtest score between reference and focal group examinees on similar ability is 0.1. The SIB procedure provides two DIF indices, β_{uni} for detecting uniform DIF and β_{cro} for detecting non-uniform DIF. For uniform DIF, the hypothesis of interest is

$$H_0: \beta_{uni} = 0 \quad vs. \quad H_a: |\beta_{uni}| > 0$$

For non-uniform DIF, the hypothesis of interest is

$$H_0: \beta_{cro} = 0 \quad vs. \quad H_a: |\beta_{cro}| > 0$$

Let $X = \sum_{i=1}^{n} U_i$ be the total score on the valid subtest and $Y = \sum_{i=n+1}^{N} U_i$ be the total score on the studied subtest. Let $\bar{Y}_{Rk}$ and $\bar{Y}_{Fk}$ be the average score in the "suspect" subtest for all examinees in the reference and the focal groups respectively attaining a "valid" subtest score $X = k$, $(k = 0, 1, 2, \ldots, n)$. Since $(\bar{Y}_{Rk} - \bar{Y}_{Fk})$ is the difference in performance in the suspect subtest across the two groups among examinees of the same ability, $(\bar{Y}_{Rk} - \bar{Y}_{Fk})$ will be expected to be equal to zero if the suspect subtest items do not show DIF. However, when there are differences in the ability distribution of the reference and the focal groups, even in the case of no DIF, $(\bar{Y}_{Rk} - \bar{Y}_{Fk})$ is known to differ systematically from zero and will tend to indicate the presence of DIF even though no DIF is present (Shealy & Stout, 1993). Therefore, if differences in ability distribution of the reference and focal groups exist, it is necessary to effect a model based adjustment known as the regression correction on the means

of $\bar{Y}_{Rk}$ and $\bar{Y}_{Fk}$. According to Shealy and Stout (1993), with the regression correction in place, cautions about the observed score as the matching criterion in place of true scores do not apply to the SIB procedure. For more details for a classical and test theory justification of the regression correction, refer to Shealy and Stout (1993). It follows that an estimate $\hat{\beta}_{uni}$ of β_{uni} is defined as

$$\hat{\beta}_{uni} = \sum_{k=0}^{n} \hat{p}_k \, (\bar{Y}_{Rk} - \bar{Y}_{Fk}) \tag{21}$$

where $\hat{p}_k$ is the proportion among the focal group examinees attaining a score of X = k on the "valid" subtest. The SIB test statistic B_{uni} for testing the hypothesis of no uniform DIF is defined as

$$B_{uni} = \hat{\beta}_{uni} / \hat{\sigma}(\hat{\beta}_{uni}), \tag{22}$$

where $\hat{\sigma}(\hat{\beta}_{uni})$ is the estimated standard error of β_{uni}. The expression for $\hat{\sigma}(\hat{\beta}_{uni})$ is given in Shealy and Stout (1991). An estimate $\hat{\beta}_{cro}$ of β_{cro} is defined as

$$\hat{\beta}_{cro} = \sum_{k=0}^{k_o-1} \hat{p}_k \, (\bar{Y}_{Fk} - \bar{Y}_{Rk}) + \sum_{k=k_o+1}^{n} \hat{p}_k \, (\bar{Y}_{Rk} - \bar{Y}_{Fk}) \tag{23}$$

where k_o is the valid subtest score at which crossing is estimated to occur. The SIB test statistic B_{cro} for testing the hypothesis of no non-uniform DIF is defined as

$$B_{cro} = \hat{\beta}_{cro} / \hat{\sigma}(\hat{\beta}_{cro}), \tag{24}$$

The expression for $\sigma(\hat{\beta}_{cro})$ is given in Li and Stout (1993). The SIB statistics B_{uni} and B_{cro} have approximate N(0,1) distributions when no DIF is present. The null hypothesis of no DIF is rejected at level α if the value of B_{uni} or B_{cro} exceeds the upper $100(1-\alpha)$th percentile point of the standard normal distribution.

The Standardization Procedure. The Standardization (STD) procedure developed by Dorans and Kulick (1986) is a non-parametric approach that focuses on the differences between the expected and observed proportion correct scores at each score level. Examinees response data based generally on number right scores are used to group the reference and the focal group members into score levels. To obtain the most stable estimates across the score range, conditional probabilities of success at each score level are developed on the reference groups. The differences between the p-values for the two groups at each score level are computed.

The standardization is computed using a weighting function which is ideally the number of focal group members at a given score level. These weights are used for the purpose of weighting each of the individual p-differences at each score level. These weighted differences are then summed across score levels to obtain an item discrepancy index, D_{std}, between the focal group and the reference group members for each item. It is given by

$$D_{std} = \Sigma_s [P_{fs} - P_{rs}] / \Sigma K_s \qquad (25)$$

where $K_s/\Sigma K_s$ is the weighting factor at each score level s supplied by the standardization group to weight differences in performance between the focal group P_{fs} and the reference P_{rs}. A flagging

criterion of D_{std} of +/-.05 is currently being used by Educational Testing Service (ETS), as a reasonable cutoff for DIF effect size.

The D_{std} provides two indices, a signed index and an unsigned index. A signed index is not sensitive to non-uniform DIF which occurs when there are differences in the slopes of the p-values for the two groups. Dorans and Kulick developed an unsigned index which is the root mean squared deviations of the p-values for the two groups. It is given by

$$RMSWD = [\Sigma K_s \ (P_{fs} - P_{rs})^2 \ / \ \Sigma K_s]^{.5} \qquad (26)$$

The major advantage of this procedure is that like MH, it compares the probabilities of a correct response for two groups of interest of the same ability. The major disadvantage of this procedure is that it has no associated test of significance. Moreover, it requires very large sample sizes to produce stable estimates (Dorans & Kulick, 1986). Overall, this is a very promising procedure when very large sample sizes are available.

Log-linear Methods and the Logistic Regression Procedure

Mellenbergh's Logit Model. Log-linear and logit models provide methods for analyzing qualitative data within the structure of multivariate data analysis. Log-linear models are used in a manner analogous to the analysis of variance procedure using regression analysis where the dependent and independent variables are discrete. Logit models are used when the dependent variables (response variables) are dichotomous. A modification of Scheuneman's chi-square

method is provided by Mellenbergh (1982) that conforms to the general structure of log-linear and logit models for contingency tables.

In log-linear method, the two-way contingency table (score category by group) of Scheuneman's chi-square method for correct responses is modified to accommodate correct and incorrect responses in a three-way contingency table (score category by group by item response category). As in the chi-square method, the response data are summarized in score category by group by item response category. The expected cell frequency in the ith score category (i = 1, 2, 3,....., s), jth group (j = 1, 2, 3,.....,g) and kth item response category, where k = 1 for a correct response and k = 2 for an incorrect response is denoted by F_{ijk}. The equation to the logit model is defined as the natural logarithm of the ratio of correct to incorrect responses specified by score category, group and interaction between the two factors and is given by

$$\ln (F_{ij1}/F_{ij2}) = C + S_i + G_j + (SG)_{ij} \qquad \begin{matrix} i = 1,\ldots, s \\ j = 1,\ldots, g \end{matrix} \qquad (27)$$

with constraints

$$\sum_{i=1}^{s} S_i = 0, \quad \sum_{i=1}^{g} G_j = 0, \text{ and } \sum_{i=1}^{s} (SG)_{ij} = \sum_{j=1}^{g} (SG)_{ij} = 0 \qquad (28)$$

where C is the overall difficulty parameter, S_i represents the main score category effect, G_j represents he main group effect, and $(SC)_{ij}$ represents the score category by group interaction effect. F_{ij1} and F_{ij2} are the number of examinees with correct and incorrect responses respectively in score category i and group j. The basic difference between the analysis of variance model and the logit model is that in

the analysis of variance model, the dependent variable is continuous whereas in the logit model, the dependent dichotomous response variable is transformed using the natural logarithm of correct and incorrect responses.

The logit model described above is a saturated model, i.e., it perfectly describes the cell frequencies. From the saturated logit model, two other non saturated models can be derived by deleting the interaction parameter or by deleting the group as well the interaction parameter. Deletion of the interaction term will yield the following model:

$$\ln (F_{ij1} / F_{ij2}) = C + S_i + G_j \qquad (29)$$

Deletion of the group effect will yield the following model:

$$\ln (F_{ij1}/F_{ij2}) = C + S_i \qquad (30)$$

with the constraints specified for the saturated logit model.

If the data fits the model described in equation (30), then the item in question is non DIF. If the model given in equation (29) fits the data, the inclusion of the group effect parameter G_j in the model indicates that the probability of a correct response conditioned of score categories differs from group to group. Therefore the parameter G_j can be interpreted as contributing towards uniform bias. In addition to the main group effect, the inclusion of $(SG)_{ij}$ indicates the interaction between score category and group and can therefore be interpreted as contributing towards non-uniform bias.

The parameters of the logit models as well as the standard errors can be estimated from a sample using the maximum likelihood procedure. Expected cell frequencies of the unbiased model given in

equation (3) are obtained and the fit of the model is assessed by computing the Pearson chi-square statistic or the likelihood chi-square statistic. Pearson's chi-square statistic is given by

$$\chi^2 = \sum_{i=1}^{s} \sum_{j=1}^{g} \sum_{k=1}^{2} (f_{ijk} - \hat{f}_{ijk})^2 / f_{ijk} \qquad (31)$$

and the likelihood chi-square is given by

$$\chi^2 = 2 \sum_{i=1}^{s} \sum_{j=1}^{g} \sum_{k=1}^{2} [f_{ijk} \ln (f_{ijk}/\hat{f}_{ijk})] \qquad (32)$$

where f_{ijk} is the observed cell frequency and $\hat{f}_{ijk}$ is the expected cell frequency. Both statistics are distributed asymptotically as a chi-square distribution. The degrees of freedom will be appropriate to the saturated or unsaturated model chosen for the analysis. For the model in which the interaction term is excluded, the degrees of freedom is equal to (s-1). For the model that in which both the interaction and group terms are omitted, the degrees of freedom is equal to s.

Van der Flier, Mellenbergh, Ader and Wijn (1984) described an iterative procedure for using the logit model to obtain more efficiency in detecting biased items. In this procedure, for each item (item m), total test score is computed for each examine excluding the item being tested (item m). Score categories are then formed from the total test scores of all the examinees (excluding item m) and the likelihood ratio chi-square for item m is computed. Then, the item with the highest chi-square values is identified and the total scores are recalculated excluding this item as well as the item being

examined. Score categories are again formed and the procedure of identifying the item with the highest chi-square is again repeated. The above steps are continued until a prescribed number of iterations has been performed or the maximum chi-square of a set of unbiased items is less than a prespecified critical value.

The advantage of applying logit models over the other contingency-table methods is that they provide more information in the detection of uniform and non-uniform bias and also they require small sample sizes. The major disadvantage of the method over the other classical procedures is that the software available for use of the algorithm requires expertise and is also very expensive.

The Logistic Regression Procedure. The standard equation to the logistic regression model for predicting the dichotomous response variables given a set of independent variables (Bock, 1975) is represented by

$$P(u_{ij} = 1) = \frac{\exp(\beta_{0j} + \beta_{1j}\theta_{ij})}{[1 + \exp(\beta_{0j} + \beta_{1j}\theta_{ij}]}, \quad \begin{matrix} i = 1, \ldots, n; \\ j = 1, 2 \end{matrix} \qquad (33)$$

where

$P(u_{ij} = 1)$ = response of person i in group j to the item,

β_{0j} = the intercept parameter,

β_{1j} = the slope parameter for group j,

θ_{ij} = the observed "ability" of an individual i in group j.

According to the definition of differential item functioning, an item is unbiased if individuals having the same ability have different probabilities of answering an item correctly. Therefore, in the above model, if $\beta_{01} = \beta_{02}$, and $\beta_{11} = \beta_{12}$, it follows that the logistic

regression curves for the two groups are the same and the item is unbiased.

By definition, an item is uniformly DIF if the logistic regression curves for the two groups are parallel, but have different intercepts, i.e., if $\beta_{01} \neq \beta_{02}$, but $\beta_{11} = \beta_{i2}$. An item is non-uniformly DIF if there is an interaction between the ability level and group membership, i.e., $\beta_{11} \neq \beta_{12}$. In the case of non-uniformly DIF, the logistic regression curves are not parallel.

The logistic regression model can be reparametrized to include a parameter corresponding to uniformly DIF and a parameter corresponding to non-uniformly DIF in the form

$$P(u_{ij} = 1) = \frac{e^{zij}}{1 + e^{zij}} \tag{34}$$

where

$$z_{ij} = \tau_0 + \tau_1\theta_{ij} + \tau_2 g_j + \tau_3(\theta_{ij}g_j) \tag{35}$$

In the above equation,

$P(u_{ij})$ = probability of a correct response for an individual i in group j,

τ_0 = the intercept,

τ_1 = coefficient of ability,

τ_2 $(= \beta_{01} - \beta_{02})$ = group difference,

τ_3 $(= \beta_{11} - \beta_{12})$ = interaction between group and ability.

where g represents group membership so that

$$g_i = \begin{cases} .5 & \text{if } j = 1 \\ -.5 & \text{ig } j = 2 \end{cases}$$

In the above model, an item shows uniform DIF if $\tau_2 \neq 0$ and $\tau_3 = 0$, and non=uniformly DIF if $\tau_3 \neq 0$ (whether or not $\tau_2 = 0$). In the model given by equation (33), the parameters of each item can be

estimated by using the method of maximum likelihood. Estimation is carried out by maximizing the likelihood function given by

$$L(u_{ij}/\theta) = \prod_{i=1}^{N} \prod_{j=1}^{n} P(u_{ij})^{u_{ij}}[1 - P(u_{ij})]^{1-u_{ij}} \qquad (36)$$

where N is the sample size, n is the test length, $u_{ij} = 1$, and $P(u_{ij})$ is as given earlier. The estimates of the parameters obtained by the maximum likelihood procedure are normally distributed with mean vector $\underline{\tau}$ (the true values of the parameters) and variance-covariance matrix which is equal to the matrix of second derivatives if the log of the likelihood function. Thus,

$$\hat{\underline{\tau}} \sim N(\underline{\tau}, \Sigma)$$

The asymptotic standard error of the estimate of τ_s (s=1,...,4) is the square root of the s'th diagonal element of Σ, i.e.,

$$SE(\hat{\tau}_s) = [\Sigma^{ss}]^{1/2}$$

Testing the hypotheses regarding the presence of DIF in test items requires testing hypotheses about some of the elements of $\underline{\tau}$. The hypotheses of interest are therefore $\tau_2 = 0$ and $\tau_3 = 0$. In general, any hypotheses stated in the form $H_0: \tau_i = \tau_{0i}$ can be tested with the z-statistic, i.e.,

$$z = \frac{(\hat{\tau}_i - \tau_{0i})}{\sigma_{ii}} \sim N(0,1) \qquad (37)$$

where σ_{ii}, the standard error is the square root of the i'th diagonal element of the variance-covariance matrix Σ.

The Mantel-Haenszel procedure can be considered as based on the logistic regression model where the ability variable is discrete and

there is no interaction between the ability variable and the group membership variable. If there are m levels of discrete ability variables corresponding to a test of (m-1) items, then a variable x_i (i=1,....,m-1) can be defined such that

$$x_i = \begin{cases} 1 \text{ if examine is in ability group } j \\ 0 \text{ otherwise} \end{cases}$$

All examinees in ability group m receive a score of -1 (or zero). Using this coding, z_{ijk} is defined as

$$z_{ijk} = \beta_0 + \sum_{i=1}^{m-1} \beta_i x_i + \tau g_j \qquad (38)$$

The parameter τ is the difference between group 1 and group 2 after adjusting for the variables $x_1,\ldots,x_{m-1}$. With this formulation, the logistic regression model can be expressed as

$$\log \frac{P}{(1 - P)} = \beta_0 + \sum_{i=1}^{m-1} \beta_i x_i + \tau g_j \qquad (39)$$

In this above case, $\tau = \log \alpha$ where α is the Mantel-Haenszel odds ratio. Testing the hypotheses that there is no group difference, i.e., τ=0 is equivalent to testing the hypothesis that α=1.

Research Studies Related to the Procedures for Detecting DIF

A number of research studies have compared different DIF detection procedures described in the previous sections using empirical as well as simulated datasets. In studies using empirical datasets, one or more DIF procedures have been compared using real datasets. Empirical data studies have the advantage of having datasets in which the actual examinee responses are available. The

main drawback of such datasets is that they do not provide any indication of the exact nature and extent of DIF present in the data sets. On the other hand, in simulated datasets, the amount of DIF can be controlled by the researcher. But it is unlikely that the simulated datasets will reflect real response datasets in all respects. Some studies have also been undertaken combining the two strategies. In such studies, DIF is manipulated in real data sets to examine whether or not DIF detection procedures are able to detect DIF present in the items.

One of the main limitations in the comparative studies on different methods for detecting DIF was that these methods failed to take into consideration the presence of uniform and non-uniform DIF present in the studied items. Since many methods for developed for detecting DIF were not capable of detecting non-uniform DIF, there was lack of agreement between the methods in the comparative studies. Therefore, the conclusions reached in these studies may not be quite valid.

In the following sections some of the research studies undertaken to investigate and compare the efficiency of the DIF procedures are presented. Comparative studies using the procedures that are being investigated in this study and other currently popular procedures are presented first followed by studies that include the less popular procedures.

Roussos and Stout (1993) presented the results of two simulation studies that investigated the effects of small sample size and item parameters respectively on SIBTEST and Mantel-Haenszel Type I error rates. The first study simulation study investigated the Type I error

rates of their assumed distributions at the significance level of .05 for sample size of 100 examinees in the reference and focal groups. Data were generated using the three-parameter logistic model and item parameters were estimated from a 25-item ASVAB autoshop test. One item from this test (a = 1.32, b = 0.03, and c = 0.25) was chosen to investigate the Type I error rates. The reference and focal groups were sampled from a normal distribution with three levels in the mean differences (d_T) (0.0, 0.05, and 1.0) and standard deviation equal to one. Four levels of equal sample sizes (100, 200, 500, and 1000) in each group were investigated with 400 simulations each. The results of the study showed no significant difference in performance between M-H and SIBTEST with both statistics adhering quite well to the nominal level of significance for small sample sizes.

In the second simulation study, the 25 ASVAB autoshop test items were used as a valid subtest to investigate the Type I error rates at .05 level of significance for a variety of test items obtained by crossing three levels of a-parameters (0.4, 1.0, 2.5), five levels of b-parameters (-1.5, -0.5, 0.0, 0.5, 1.5). The c-parameters were set equal to 0.20 for all cases. Additionally, when the ability distribution differences were one standard deviation lower, the c-parameters were set to 0.10 and 0.05. The data were replicated 100 times for each studied item. Three levels of sample sizes 500, 1000, and 3000 were used in the study. In all 45 conditions were simulated.

For the case of d_T = 0 and 1.0, with c=.20 both procedures adhered quite well to the nominal level Type I error rates. For the case when d_T is equal to zero and c = 0.10, and c = 0.05, for sample sizes, 1000 and 3000, the Type I error rates were slightly inflated.

The results suggested that both procedures had higher Type I error rates as the sample size increased, the inflation being higher for MH than for SIBTEST. The MH and SIBTEST tend to have higher Type I error rates with respect to items with high discrimination and low difficulty. For the group size of 300, the MH also showed increased Type I error rates for items of low discrimination and high difficulty, whereas, SIBTEST had little or no inflation. Overall, SIBTEST adhered much better to the nominal 0.05 rejection rate than the MH.

Swaminathan and Rogers (1990) conducted a study with simulated data sets to investigate the detection rates of the Mantel-Haenszel and logistic regression procedures. In this study, two levels of sample sizes (250,500), and three levels of test lengths (40, 60, 80) were crossed to produce six conditions. Within each test, 20% of the items, half uniformly DIF and the other half non-uniformly DIF were investigated for the capability of both procedures to detect DIF items.

Item responses for the study were generated with specified item parameters using the three-parameter model. For uniformly DIF items, the b differences were chosen (.48 and .64) so as to obtain a prespecified area (.6 and .8) between the ICCs for the two groups. The b difference values so chosen represented moderate to high bias. For non-uniform DIF items, the item discrimination values were chosen so as to obtain the areas (.48 and .64) between the ICCs for the two groups.

The results indicated that the detection rates for the two procedures were similar for uniformly DIF items. Both methods were

able to detect DIF items with about 75% accuracy in small sample sizes and short tests and with about 100% accuracy in larger sample sizes and longer tests. The false positive rate was around 1% for the Mantel-Haenszel procedure and was between 1% to 6% for the logistic regression procedure,at a significance level of .01, indicating that the performance of the Mantel-Haenszel procedure was slightly better than the logistic regression procedure.

The results also indicate that for a test length of 80 items and sample size of 500 examinees, over 20 replications, the detection rate for the Mantel-Haenszel procedure was 96% for uniform bias, 1% for non-uniform bias and a 1% false positive error rate. For the logistic regression procedure, the detection rate was 94% for uniform bias, 71% for non-uniform bias with 4% false positive errors.

The authors concluded that the logistic regression procedure was as powerful as the Mantel-Haenszel procedure in detection uniform bias and more powerful than the Mantel-Haenszel procedure in detecting non-uniform bias.

Rogers (1989) conducted a simulation study comparing the MH and the logistic regression (LR) procedures. The first part of the study investigated the distributional properties of the MH and the LR procedure. Two conditions were examined for the distributional properties of the LR procedure to determine the conditions under which (1) the estimates of the logistic regression procedure were distributed normally and (2) the LR test statistic was distributed as a chi-square with two degrees freedom. The MH statistics was examined for a chi-square distribution with one degree of freedom.

The factors manipulated in this phase of the study were two levels of the degree of model fit (using the 2P-IRT and the 3P-IRT models) and two levels of sample size (250, 500) in each group. Data for the study was simulated according to the two-parameter logistic model for a 45-item test in which all items were non-DIF. In the 45-item test, five items with different combinations of difficulty and discrimination parameters (low, medium, high) were studied for distributional properties with 100 replications of each data set.

The results show that the estimates of the logistic regression procedure were in general, normally distributed as expected without any marked effects due to sample size and model-data fit. On the whole, the LR test statistic was seen to be distributed as a chi-square under all conditions except one. Investigation of the MH statistic showed that it was not distributed as a chi-square for a larger number of cases. Sample size did not seem to have an impact on the distributions. The Type I error rates at the 95th and 99th percentile cut-off points appeared to be acceptable for both procedures under most conditions. The LR procedure had slightly higher false positive rates than the MH procedure. The LR and MH seemed to conform well to underlying theory and hence at this stage appeared to have the potential to be a useful indicator of the presence of DIF.

The factors manipulated in the power study were two levels of the degree of model-data fit (using the 2P-IRT and the 3P-IRT models), two levels of sample size (250, 500) in each group, two levels of test length (40, 80), two levels of proportion of DIF items (0%, 15%), two levels of test score distributions (normal, skewed), four levels of

DIF effect size (.2, .4, .6, .8). In all, 256 conditions were simulated.

The study was conducted to investigate uniform and non-uniform DIF. Five different combinations of the difficult and discrimination parameters (low, medium, high) for each of the four DIF effect sizes under investigation were studied for non-uniform DIF for a total of 25 items. Four combinations of the difficulty and discrimination parameters for a total of 20 items were studied for uniform DIF. In all, 35 items representing various levels of difficulty and discrimination parameters were studied for DIF for data simulated according to the two and three-parameter logistic models. Twenty replications of each data set were carried out.

The results show that the detection rates for all procedures increased for increase in sample size, decrease in the proportion of items containing DIF, and increase in DIF effect size. Detection rates were highest for items of medium difficulty and high discrimination, and lowest for items of medium difficulty and low discrimination.

The detection rates when DIF was uniform were almost similar for the MH and LR procedures with the MH procedure performing slightly better. In the case of non-uniform DIF, the identification rates for the LR procedure was much better than the MH procedure. Detection rates for the LR procedure varied between 35% and 85% with most cases between 50% and 70%. The detection rates for the MH procedure were between 5% and 80% with most cases. The detection rates for the LR procedure increased with better model-fit to the data, and increasing sample size, increasing DIF effect size. The results show that DIF

items of medium difficulty and high discrimination were most likely to be identified and DIF items of medium difficulty and low discrimination were least likely to be identified.

The author concluded that, while the MH and the LR procedures were equally effective in detecting uniform DIF, the LR procedure was much more effective than the MH procedure in detecting non-uniform DIF.

Hambleton and Rogers (1989) conducted a real data study comparing the IRT based area method and the Mantel-Haenszel procedures for investigation DIF. This research investigated the degree of agreement between the two methods in identifying DIF and examined the consistency with which each DIF statistic identified biased items when the ability distributions of the two groups of interest were different.

The data for the study consisted of responses of 2000 Anglo American and 2000 Native American students drawn from a data set containing responses of approximately 23,000 students to the New Mexico High School Proficiency Exam (NMHSPE). Two samples of 1000 students were obtained by randomly sampling the 2000 Anglo American students (Sample 1 and Sample 2). Similarly, two samples of 1000 students were obtained by randomly sampling 1000 Native American students (Sample 1 and Sample 2). Out of the total of 150 items in the original test, 75 items were chosen for the study.

In the area method, the area between the ICCs for the two groups was computed and the DIF statistic values were computed. A cutoff value of the area statistic was obtained by carrying out an analysis on two randomly equivalent groups (the two Native American samples).

The largest area statistic obtained (.468) was used as an indicator of the greatest value of the statistic that can occur by chance. Since the sampling distribution of the area statistic is not known, the items were ranked according to the DIF statistic values and the items with the highest values were identified as DIF. The cutoff value used for the Mantel-Haenszel statistic was 6.64 which is the tabulated value of the chi-square distribution with 1 degree of freedom at the .01 alpha level.

The results of the investigation of the power of the two procedures showed that both methods were somewhat unstable across samples. The overall consistency of the Mantel-Haenszel and the area methods were 80% and 73%, respectively. For the area method, 61% of the items flagged in Sample 1 were flagged by Sample 2 while for the MH procedure, 47% flagged in Sample 1 were flagged by Sample 2. On the other hand, 56% and 64% of the items in Sample 2 were flagged in Sample 1 for the area method and the Mantel-Haenszel method. Out of the 14 items consistently identified by the area method across the samples across two comparisons and the nine items consistently identified by the Mantel-Haenszel method, seven items were common. Several of the discrepant items were explained in terms of Type I errors and several of the items explained as being due to the presence of non-uniform bias, which the Mantel-Haenszel method was unable to detect.

To study the effect of the score distribution differences on the Mantel-Haenszel statistic, a matched-group analysis was carried out by selecting a third sample of Native Americans such that the distribution of scores closely matched that of the Anglo American

sample. Results from the matched group comparison showed very little change for the Mantel-Haenszel statistic results, whereas the area method results showed greater change, although this may have been due in part to the sample-size reduction that was necessary to achieve matching. The authors concluded that the Mantel-Haenszel method can be safely substituted for IRT-based methods if safeguards suggested by the authors are put in place to detect nonuniform bias.

Wise (1987) conducted a simulation study comparing nine DIF procedures which included the transformed item difficulty method, the Camilli signed and unsigned chi-square procedures, two modifications of the delta method, the MH chi-square and the MH-delta, a signed version of the MH procedure and the compound binomial exact test.

The study examined 96 conditions obtained by crossing two levels of sample size (400, 800), two levels of test length (30, 60), two levels of proportion of DIF items (10, 20), two levels of mean difference in the ability distribution between the two groups (0, 1 s.d.) and, three levels of the ratio of focal group standard deviation to reference group standard deviation (0.5, 1, 2).

Data for the study were simulated using the three-parameter logistic model for different combinations of the a, b and c parameters. The DIF effect sizes were simulated by specifying three levels in b-parameter differences (.1, -.4, .8). The results were evaluated based on a cut-off point representing the 95th percentile of the distribution of the DIF statistics obtained for all non-DIF items.

The results averaged across all the three DIF effect sizes showed that, over all conditions, the signed MH chi-square statistic performed the best followed by the signed chi-square the MH chi-square

statistics. The transformed item difficulty performed better than the unsigned chi-square procedure. The detection rates for all procedures ranged from 4% to 10% for DIF effect size equal to .1, from 25% to 40 % for DIF effect size equal to -.4 and from 65% to 80% for DIF effect size equal to .8. The detection rates were higher when the discrimination was moderately high (1.5) and when the guessing parameters were lower.

The identification for all procedures increased for increase in sample size, for increase in proportion of DIF items and for equal ability distributions. Test length did not appear to affect any of the procedures. Wise concluded that the DIF procedures investigated in this study were useful in detecting DIF in small sizes,

Wright (1986) compared the performance of the Mantel-Haenszel and the standardization procedures with samples drawn from a 85-item verbal section of the November 1984 administration of Scholastic Aptitude Test. A "full" sample of 10,000 Whites (reference group) and 3000 Blacks (focal group) formed the basis for the selection of five subsamples for the study. The subsamples consisted of different combinations of reference and focal groups: (3000 and 3000), (10,000 and 800), (5000 and 400), (2000 and 200), and (1000 and 80). On the SAT scale of 200-600, two levels of score groups, one using a score interval width of 10-points producing 61 score groups, and the other using a score interval width of 100-points producing six score groups were used as the matching criteria to compute the statistics.

The results were presented in terms of means, standard deviations and correlations between two MH statistics, α_{MH} and Δ_{MH}, and three standardization indices, D_{STD}, α_{STD}, and Δ_{STD} for the "full" sample

and two levels of score groups. It was seen that the correlations between Δ_{MH} and Δ_{STD} was .99 for both 10-point and 100-point intervals for the "full sample". The means, standard deviations and correlations of Δ_{MH}, D_{STD}, and Δ_{STD} were also presented across each sample size and the two levels of score groups. The results indicated that the correlations ranged from .98 to .99 for large sample sizes, and ranged from .93 to .98 for the smallest sample size. The number of score intervals had little effect in the ordering of the items. However, there was a constant difference in means between 10-point and 100-point intervals. The results also show that the D_{STD} was the most stable index followed by Δ_{STD} and Δ_{MH}. Wright concludes that "six standardized intervals are inadequate for matching such extreme groups (mean differences of approximately .8 standard deviations)" (p. 9).

The standardized differences D_{STD} were collapsed into three intervals corresponding to < -.05, -.049 to .049, and > .05 and three replicated sample sizes (5000 and 400), (2500 and 200) and (1000 and 80) were crosstabulated against the full sample. The percentages of agreement ranged from about 80% for sample size (5000 and 400), about 70% for sample size (2500 and 200) and about 55% - 60% for sample size (1000 and 80).

The results of the study show that the standardized difference and the MH common odds ratio statistics measure very similar phenomenon for the data under study. It appeared that the STD difference statistic is slightly less subject to sampling fluctuations, but if both statistics were transferred to the ETS delta scale, the MH delta showed slightly more stability.

Shepard, Camilli and Williams (1984) conducted a real data study comparing several three-parameter IRT techniques to study statistical artifacts associated with IRT DIF indices. They were: (1) Unsigned area (UA) - the absolute value of the area between the ICCs for the two groups, (2) Sum of squares 1 (SOS1) - sum of the squared differences between the ICCs for the two groups at each ability level, (3) Sum of squares 2 (SOS2) - sum of the squared differences weighted by the inverse of the variance error of the difference in ICCs at each ability level, (4) chi-square (IRT χ^2) based on Lord's significance test comparing a and b differences simultaneously, (5) Signed area (SA) - same as UA with a positive and a negative sign attached if DIF was against the focal and the reference group respectively, (6) Sum of squares (SOS3) - the signed sum of squared differences between the ICCs for the two groups at each ability level, and (7) Sum of squares (SOS4) - the signed weighted sum parallel to SOS2.

The data sets used in the study consisted of subsamples of black and white students drawn randomly from the High School and Beyond (HSB) data files on a mathematics as well as a vocabulary test. The study samples created for the math test were: (1) Comparison 1: (W1B1) - 1500 whites, 1500 blacks; (2) Comparison 2: (W2B2) - 1500 whites, 1500 blacks; (3) Comparison 3: (W1W2) - white samples from (W1B1) and (W2B2); (4) Comparison 4: (B1B2) - black samples from (W1B1) and (W2B2); and (5) Comparison 5: (W1W3) - white sample from (W1B1) and a white sample (1500) selected to match the distribution of B1 on math total score. The study sample created for the vocabulary are: (1) Comparison 1: (W4B4) - 1500 whites, 1500 blacks; (2) Comparison 2: (W5B5) - 1500 whites, 1500 blacks; and (3) Comparison 3: (W4W5) -

white samples from (W1W3). There were a total of 32 items in the math test and 29 items in the vocabulary test.

The seven DIF indices were computed separately for the math and the vocabulary tests using the LOGIST program to estimate the item and ability parameters for the two groups. Two randomly selected groups of whites (W1B1) and (W2B2), were used to obtain a baseline for providing a cutoff value for flagging items as DIF. The results for (W1B1) and (W2B2) indicated that 10 of the 29 items for which the ICCs could be estimated for the two groups were consistently detected to be DIF of which three items were favoring the blacks. (W1W2) and (B1B2) being basically comparisons between two randomly equivalent groups, the results showed that the DIF indices were appreciably smaller than in white-black comparisons. The DIF indices (both signed and unsigned) were substantially smaller in the (W1B1) than in (W2B2). In (W1W3), with the exception of IRT Δ^2, no DIF was detected by the indices, indicating that large differences in the black-white comparisons might be due to real differences in the functioning of items across groups. Results for the vocabulary test corroborated the findings based on the math test, but verbal math problems were found to be systematically biased against blacks.

To examine the relationships between the DIF indices, within-study and between-study Spearman rank order correlations were obtained for the DIF indices for each comparison in the study. The results indicated that the correlations were not high for the items that were identified as functioning differently. The signed indices were less correlated than the unsigned indices. The pattern of between-study correlations showed high consistency between analyses where DIF was

present. Also the correlations were low between conditions where bias should not be present.

The results indicated that the signed sum of squares showed the greatest agreement (90%) over the two replications in identifying the items as biased or unbiased. Lord's chi-square method showed 86% agreement, and the area method showed 83% agreement. The signed indices were less correlated than the unsigned indices. The authors concluded that the sum of squares indices (SOS2, SOS3, SOS4) were most consistent for detecting DIF based on IRT methods. There statistics were not only most consistent in detecting DIF in the ethnic groups, but they also intercorrelated the least in situations of no bias.

Shepard, Camilli and Williams (1985) conducted a study with real as well as simulated data to investigate the validity of the most popular approximation techniques for detecting DIF. The methods tested in the study were: (1) the transformed item difficulty (TID), (2) the three-parameter IRT which included seven different bias indices, (3) Camilli's chi-square, and (4) the pseudo-IRT.

The seven bias indices investigated in the IRT method were the unsigned (UA) between the ICCs for the two groups, the sum of squared differences (SOS1) between the ICCs, the sum of the squared differences weighted by the inverse of the variance error in the ICCs, Lord's chi-square index (IRT χ^2), the signed area (SA) between the ICCs for the two groups, the signed sum of squared differences (SOS3) between the ICCs and the signed sum of squared differences (SOS4) weighted by the inverse of the variance error in the ICCs were computed.

The data for the study consisting of two groups of 1000 whites and 300 blacks was randomly subsampled from the data for two groups of 1500 whites and 1500 blacks used in their previous study (Shepard et al., 1984). The total number of items in the math test was 32.

The results based on the baseline values obtained for interpreting the magnitude of the DIF indices showed that 10 items were consistently found to function differentially with seven items favoring the whites. The results of the signed indices showed that, the weighted SOS4 were judged the most valid. The pseudo-IRT procedure correlated best with the SOS4 indices (.72). The chi-square technique had the next best correlations to the IRT indices (.65). The TID correlated least well with the IRT indices. False positive rates were fairly low for pseudo-IRT (3 out of 10), and chi-square methods (2 out of 10). In terms of actual detection rate, the signed chi-square and pseudo-IRT correctly classified 90% of the items as differentially functioning.

The study also investigated the validity of the DIF techniques with data generated according to the three-parameter IRT model for the 300 black examinees and 1000 white examinees. Ability parameters were set to have a mean of 0.0 for the 300 blacks and a mean of 0.8 for the 1000 whites. The standard deviations for both the groups were set to 1.0. Fifty-four items in nine cells were generated for the white group with six items in each cell and the item parameters were obtained by crossing three levels of discrimination (0.5, 1.0, 1.5) with three levels of difficulty (-1.0, 0.0, 1.0). The c-parameters for all the items were set equal to 0.25. For the black group, 36 items were generated in the same manner as before with four items in

each of the nine cells. Out of the remaining 18 items, bias was introduced by increasing the b-parameter value by 0.20 for 9 items ("weak' bias), and by increasing the b-parameter value by 0.35 for the remaining 9 items ("moderate bias").

The results of the bias indices showed that the signed indices had generally higher correlations than the unsigned bias indices. The signed sum of squares differences SOS3 (.61), SOS4 (.61) and the signed pseudo-IRT indices (.62) had high correlations. The correlation for signed chi-square statistic was .59 and the transformed difficulty was .46. In terms of the actual detection rate, the signed chi-square and pseudo-IRT index were able to correctly classify 74% and 76% respectively,of the items as functioning differently when all the items were included. When only moderately biased items were included the detection rate was 87% and 91% respectively. False positive rates were fairly low for the pseudo-IRT (1 out of 36 unbiased items) and the chi-square (3 out of 36), but high for the transformed item difficulty method (9 out of 36).

Shepard et al. (1985) concluded that the pseudo-IRT approach is a promising technique for use with small samples. The Camilli chi-square index was close in accuracy to the pseudo-IRT index. The TID method was found to be inadequate.

Ironson, Homan, Willis and Signer (1984) conducted a real data study to investigate the validity of three differentially functioning procedures which included, (1) the transformed item difficulty procedure, (2) the Camilli chi-square procedure, and (3) a modified

version of the pseudo-IRT procedure developed by Linn and Harnisch (1981).

The data for the study consisted of responses from a sample of 1064 students from the second and fourth grade levels. The scores from the reading and math tests of the Comprehensive Tests and Basic Skills (CTBS, 1973) were obtained separately for the total sample as an independent assessment of the students' reading and math levels. The reference group consisted of 916 students whose reading and math levels were above fourth grade. The focal group consisted of 148 students whose reading level was below fourth grade and math level was above fourth grade. A mathematical word problem test consisting of three types of items: (1) ten items at second grade reading and second grade math level, (2) ten items at second grade reading and fourth grade math level, and (3) six items at fourth grade reading and fourth grade math level, was administered to the students. The six items at fourth grade reading and fourth grade math level in the test were tested to determine if they functioned differently for the two groups.

DIF analyses were conducted based on the rank order correlation between each signed bias index and a (0 = unbiased, 1 = biased) classification of items (1 = fourth grade reading, fourth grade math item, 0 = all other items). The results showed a non significant correlation (-0.23) between transformed item difficulty method and the 0/1 classification. Also, none of the items appeared to function differently as the absolute values of the individual item distances were all < 1.0. The correlation obtained for the Camilli chi-square method was also non significant (-0.23) and none of the six items tested appeared to be functioning differently. The standardized

difference scores obtained for the modified form of pseudo IRT procedure correlated very well (.63) with the DIF classification. Five of the six items tested had DIF indices among the seven highest. The authors concluded that the IRT method was the best among the three procedures for DIF.

The validity of the test was measured by the correlation of the math word problem with the CTBS math assessment test. The correlations of the total test scores for the total sample decreased from 0.661 to 0.656 when the DIF items were removed. The validity of the test for the reference group also slightly decreased from 0.555 to 0.536 . However, removing the DIF items increased the validity of the focal group from 0.250 to 0.283. The authors concluded that DIF items had a minor effect on the validity for the focal group.

Subkoviak, Mack, Ironson and Craig (1984) conducted a study using real data with manipulated bias comparing four DIF procedures: (1) the signed and unsigned transformed difficulty procedure, (2) the Scheunaman (CHIS) procedure, (3) the signed and unsigned Camilli (CHIC) chi-square procedure, and (4) the signed and unsigned three parameter IRT procedure using the area method.

A 50-item vocabulary test consisting of 40 items coded (0) involving standard English vocabulary, and 10 items consisting of black slang words coded (1) to favor black students was constructed. The test was administered to 1008 black students (from an urban eastern university) and 1021 white College students (from a midwestern university) so as to have two groups that differed geographically as well as racially. For each item, DIF indices based on the student's responses were computed for each of the DIF procedures.

The results show that for all methods, the correlations between the apriori bias (0/1) of the items and the corresponding DIF indices involving the signed bias measures were larger than the corresponding correlations based on unsigned measures. From the results, it was also seen that the three-parameter IRT procedure was most effective at detecting apriori bias with correlations of 0.872 (unsigned) and 0.875 (signed). The correlations of the unsigned CHIC, CHIS and TID methods were 0.719, 0.731 and 0.733 respectively. The correlations of the signed CHIC, CHIS, and TID methods were 0.870, 0.799 and 0.870 respectively. The TID and CHIC methods produced very similar results for the signed and unsigned measures. Correlations among all the methods showed a high degree of relationship between them.

Intercorrelations among the DIF indices of the three DIF procedures indicated that the ICC-3 procedure correlated highest 0.883, 0.881 and 0.853 respectively with the CHIS and CHIS procedures in the analysis involving unsigned measures. In the analysis of signed measures, the ICC-3 procedure correlated about equally well with CHIC (0.933), CHIS (0.928), and TID (0.939).

The study indicated that the TID procedure is less effective than the chi-square methods because it is primarily sensitive only to bias related to item difficulty parameters. The authors concluded that the three-parameter IRT method proved most effective at detecting the apriori bias present in the item set, but requires large sample sizes and computer facilities. On the other hand, the Camilli chi-square procedure came out in the study as the next best procedure and should be used as a method of second choice.

Hoover and Kolen (1984) conducted a study with real data to examine the reliabilities of six item differential item functioning indices which included the transformed item difficulty and delta indices, biserial and point biserial indices, the Scheuneman's chi-square index and a modification of the pseudo-IRT index proposed by Linn and Harnisch (1981).

The data for the study consisted of 200 randomly selected samples each of Black males, Black females, White males, White females from responses of 800 fifth grade students from eleven subtests to the Iowa Test of Basic Skills (ITBS). Each sample of 200 Black and White students were randomly divided into two sample sizes of 200 students each to include 100 males and 100 females in each sample. DIF indices were computed for the two independent Black - White samples and the two male - female samples. DIF indices were also computed separately for each of the eleven ITBS subtests.

The investigation of the results of the study included computing the correlations of the six DIF indices obtained from the two analyses for each subtest to assess the reliability of each of the six DIF indices. Items with difficulty or delta indices above .75 or Scheuneman index values which exceeded the .05 critical value of a chi-square distribution with 4 degrees of freedom were classified as DIF. The results were also investigated for classification consistency of the DIF indices to be able to classify items as either DIF or non DIF across two comparisons, by specifying suitable cut-off values. The agreement of classifications across samples was evaluated using chi-square tests of independence with Yates' correction. To estimate the parameters of the three-parameter IRT model, the entire

sample of 800 examinees was used to obtain convergence in parameter estimates. In view of the possibility of over estimation of reliabilities, DIF indices were calculated for only two subtests (vocabulary and language usage) and other analyses were excluded.

The main results of the study indicate that reliabilities of all the DIF indices were generally very low. The reliability of the DIF indices by race ranged from -.01 to 0.55 across all subtests for the transformed item difficulty method, 0.04 to 0.55 for the Scheuneman chi-square procedure and .25 to 0.36 for the three-parameter IRT model. For the sex comparison, the reliabilities of the DIF indices were even lower than those obtained for the race comparison. The reliability of transformed difficulty procedure ranged from -0.16 to 0.22, for the Scheuneman chi-square method it ranged from -0.02 to 0.38 and for the three-parameter method it ranged from -.16 to .09.

For the race comparison, the classification consistency statistics to investigate agreement across randomly equivalent samples were 4.70 and 3.32 respectively for the transformed difficulty and delta indices. For the Scheuneman method, the value was 0.63 indicating that there was not a significant relationship. For the sex comparison, values of the classification consistency statistics were 7.86 for difficulty, 10.72 for delta and 2.09 for Scheuneman chi-square method, with the tests of difficulty and delta exceeding the critical value. The results showed that the agreement across randomly equivalent samples were not sufficiently significant, possibly due to the fact that there was very little bias in the ITBS tests due to screening of the tests by experts. The authors point that if such is

the case, the use of bias indices could not provide enough information about bias.

The results of the study indicated that reliability decisions may not be feasible using DIF indices. The authors suggest that use of simulated data sets specifying and crossing with a variety of conditions mainly to include sample sizes to assess the reliability of DIF indices especially with respect to sample sizes for further investigations.

Shepard et al. (1981) conducted a real data study to comparing six different DIF procedures. They are: (1) transformed item difficulty (TID), (2) item discrimination or point biserial (PB) in which an unsigned bias index was computed using the absolute differences between the point-biserial item total score correlations between the two groups, (3) three-parameter ICC (ICC-3), (4) one-parameter ICC (ICC-1), (5) Scheuneman's chi-square, and (6) Camilli's chi-square.

In ICC-3 method, five DIF indices were obtained from the item parameters estimates using LOGIST. They were: (1) differences in difficulty parameters, (2) differences in discrimination parameters, (3) the signed and (4) the unsigned area between the curves for the two groups, (5) a composite significance test of differences in both the a and b parameters (Lord, 1980). In the ICC-1 method, five DIF indices were computed from the item parameter estimates using LOGIST. They were: (1) the item difficulty parameters, (2) the weighted difference in difficulty parameters taking the variance of the b's constant, (3) the absolute value of the differences in the mean

squared standardized residuals for the two groups, (4) the signed and (5) the unsigned area between the two curves.

Data for the study was obtained from 490 Black, 551 Chicano, and 552 White students in the fourth, fifth and the sixth grades from an administration of the Lorge-Thorndike test, a mental ability test containing 90 verbal and 79 non-verbal items. Separate analyses were made using an internal criterion (total test score) and an external criterion in which scores on the Lorge-Thorndike test was used as the criterion for groups examined for the analyses on Raven's Coloured Progressive Matrices, a culture-fair test.

Comparison of a total of 16 DIF indices showed that the correlations between the signed full chi-square procedures and the ICC-3 area, the TID procedure and the ICC-3 area, the TID procedure and the chi-square procedure were all moderate. The unsigned indices were uncorrelated with the signed indices as well as with other indices. The TID procedure also correlated very highly with ICC-1.

Using an internal criterion, Shepard et al. found that moderate to high correlation between signed full chi-square and signed IRT-3 area indices (.68) and between Camilli's chi-square and the transformed difficulty method (.67). Correlations between signed measures were generally lower. In general, the results showed that more items were identified as DIF by using the external criterion. Shepard et al. concluded that the relationships among the different methods were strong enough for the use of a simpler method such as the chi-square as an approximation to IRT methods for detecting DIF.

Rudner, Getson, and Knight (1980b) conducted a study with simulated data to investigate seven DIF procedures. They are: (1)

transformed item difficulty procedure (TID-MA) which involved computing the absolute values of the distances of the delta points from the major axis of the line of best fit, (2) the modified transformed difficulty procedure (TID-45) which involved computing the absolute distances of two sets of p-values transformed to within group z-values from the major axis of the line of best fit, (3) the chi-square technique (CHI-5) of Scheuneman (1975) using five score group intervals for the two groups to compute the chi-square values, (4) the chi-square technique with multiple intervals (CHI-N) using the total score group intervals minus the number of cells with expected values less than five to compute the chi-square values, (5) IRT method with one parameter (ICC-IF) which involved computing the absolute value of the differences in the mean square fit of items obtained from item difficulty parameter estimates using the Rasch model, (5) IRT method with one parameter (IRT-IE) which involved computing the absolute value of the differences in the item difficulty values estimated by Rasch model, and (7) IRT method with the three parameter model (IRT-3) in which the area between the ICCs for the two groups were computed with estimated item parameters.

Data for the study were generated using the three-parameter IRT model for seven levels of test lengths (20, 30, 40,..., 80), four levels of amount of bias in difficulty parameter (0.0, 0.5, 1.0, 1.5) and four levels of amount of bias in the discrimination parameter (0.0, 0.2, 0.4, 0.8, 1.2) were crossed to produce 112 different test conditions. Two groups of 1200 examinees with one standard deviation difference in mean level of performance were also selected.

The correlations between generated and detected amount of bias across all test lengths and over all the test conditions were found to be the most accurate for ICC-3 (.80), for chi-square method with 5 intervals (.73) and for the transformed difficulty procedure (.68) and least effective for the Rasch models ICC-IF and ICC-IE.

In terms of the difficulty levels, the correlations between generated and detected amounts of bias increased for ICC-3, TID-45 and ICC-IE with increase in difficulties while the correlations for ICC-IF showed a steady decline. CHI-5 and CHI-N techniques showed a steady decline in correlations except for the extreme condition of difficulty equal to 1.5. In terms of the discrimination levels, the correlations between detected and generated bias steadily decreased with increasing difficulty except for the correlation of ICC-IF technique which steadily increased and the ICC-3 technique remained relatively stable.

Rudner and associates concluded that three of the investigated procedures, ICC-3, CHI-5, and TID-MA produced fairly accurate estimates of generated bias. Increasing the number of intervals decreased the accuracy of CHI-5 and CHI-N. The IRT approach with one parameter consistently correlated poorly with the amount of generated bias.

Summary

Procedures for detecting DIF are used to identify whether the individual items in a test function in the same way between different subgroups. Research in this area for many years have yielded a number of statistical methods for detecting DIF. These procedures can be

classified as classical test theory approaches (CTT), item response theory approaches (IRT), and Chi-square approaches.

A number of methods for detecting DIF derived from the principles of CTT include the analysis of variance method, factor analytic method, item discrimination procedure, partial correlation method, and the transformed difficulty method. The most popular CTT approaches are based on the transformed item difficulty (TID) procedures. The most widely used of the TID procedures is Angoff's delta plot method. The advantages of the delta plot method are that it is simple, inexpensive and does not require large sample sizes. CTT based approaches for detecting DIF make use of the observed response data for persons and items rather than the "true scores". Therefore, these procedures are to some extent, sample dependent. Because of this problem, the results from a DIF study using classical methods cannot be generalized to larger populations of interest based on the samples drawn from the groups of interest. In general, CTT based procedures for detecting DIF are not likely to be very useful to major testing programs (Hambleton, Clauser, Mazor & Jones, 1993).

IRT based approaches for detecting DIF tend to overcome the limitations of CTT approaches because of their theoretical underpinnings. The feature of item parameter invariance of IRT estimates for different samples drawn from the same population makes it specifically useful for investing the presence of DIF in test items. When an IRT model fits the data, IRT based procedures offer the advantage of using the true ability estimates rather than the observed scores. In IRT, an item is differentially functioning if the ICCs for the two groups are not identical. A number of IRT based

methods developed for detecting DIF focus on three major approaches. They are: (1) comparison of ICCs for the two groups, (2) comparison of the vectors of the item parameters, and, (3) comparison of the fit of the IRT models to the data. The main disadvantages of these methods are that they are complex, and require large sample sizes for item parameter and ability parameter estimation. Moreover, the most popular IRT methods do not have associated tests of significance.

A number of methods for detecting DIF are based on the construction of two-way contingency tables (group by item response) and offer a chi-square value as an index of DIF. Unlike CTT procedures which focus on a single item parameter such as item difficulty or discrimination in DIF analyses, these procedures compare the entire distributions of responses for the two groups of interest. In these methods, the observed scores are used to match the reference and the focal group examinees before investigating DIF.

The chi-square approaches for detecting DIF are based on the definition that an item functions differently for two groups if the probability of a correct response is not the same for persons of equal abilities in the two groups. These methods may be thought of as approximations to IRT procedures and are recommended where IRT procedures are not feasible. These procedures have several advantages over the procedures based on CTT and IRT. Besides requiring smaller sizes, these procedures will reflect the interaction effects between group differences and ability level, have associated tests of significance, and are inexpensive to use (Scheuneman, 1989).

The log-linear procedures for detecting DIF are extensions of the traditional chi-square methods and are based on the construction

of three-way contingency tables (score category by group by item response). An advantage of the log-linear models over the chi-square methods is that they are able to make a distinction between group differences (uniform DIF) as well as interaction effects between group differences and ability (non-uniform DIF). Like chi-square methods, these procedures can be used with relatively small sample sizes.

In recent years, a number of non-parametric approaches for detecting DIF have been developed as alternatives to the more complex IRT procedures. Previous research shown that, unlike IRT procedures, these procedures are computationally simple, inexpensive with respect to computer time, and effective with small sizes. Prominent among these methods are the MH, the standardization (STD) and the SIB procedures.

In recent years, the MH procedure (Holland & Thayer, 1988) has been one of the most popular and widely used procedure for detecting DIF in test items. Previous research have shown that although the MH procedure is very effective in detecting uniform DIF, it may not be sensitive to non-uniform DIF (Swaminathan & Rogers, 1990). Both the MH and the STD procedures compare the probabilities of a correct response for two groups of interest of the same ability. While the MH procedure has an associated test of statistical significance, the STD procedure provides only an index as a measure of DIF effect size. Moreover, the STD procedure requires large sample sizes to produce stable estimates (Dorans & Kulick, 1986).

The LR procedure, introduced by Swaminathan and Rogers (1990) based on the logistic regression model is becoming increasingly popular in DIF studies. The LR procedure has been found to be more

powerful than the MH procedure for detecting non-uniform DIF and as powerful as the MH procedure in detecting uniform DIF. Despite its parametric nature, the LR procedure can be easily implemented in practice.

The SIB procedure developed by Shealy and Stout (1993) emphasizes the examination of DIF at the test level. It provides a statistical test of significance that can detect DIF present in one or more items in a test simultaneously. Previous research have shown that the SIB procedure is as effective as the MH procedure in detecting uniform DIF. A modification of the SIB procedure to detect non-uniform DIF was presented by Li and Stout (1993). Because of its newness, the SIB procedure for detecting non-uniform DIF has not been extensively studied.

The detection of DIF in test items being an issue of major concern in educational data, practitioners and test developers are interested and obligated to do DIF studies. Although a variety of DIF detection techniques are currently available, no single method can be guaranteed to identify all of the DIF items in a test. However, research studies focused on the comparison of multiple methods can address the advantages and shortcomings found in a particular method. Therefore, with a variety of DIF detecting procedures currently available, empirical research to compare multiple methods is necessary to determine the conditions under which each procedure is optimal for detecting DIF.

C H A P T E R III

PERFORMANCE OF THE MANTEL-HAENSZEL AND SIMULTANEOUS ITEM BIAS PROCEDURES FOR DETECTING DIFFERENTIAL ITEM FUNCTIONING

Introduction

In recent years, the concern over the issue of differential item functioning (DIF) in standardized achievement and ability tests has resulted in the development of a variety of statistical methods for detecting DIF. The most theoretically sound procedures are based on item response theory (IRT). However, these procedures require large sample sizes, a condition that is often difficult to meet in practice in most DIF studies. Because of this problem, measurement specialists have been involved in developing non-parametric methods as alternatives to IRT procedures. The advantages of these procedures are the fact that they are effective with small sample sizes, computationally non-intensive, and cost effective.

The focus of this study was on two non-parametric procedures for detecting DIF: the Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988) and the Simultaneous Item Bias (SIBTEST, henceforth referred to as SIB) procedure (Shealy & Stout, 1991, 1993). The purpose of this study was to compare the MH and the SIB procedures to determine the conditions under which each procedure was optimal for detecting uniform DIF.

The Mantel-Haenszel and the Simultaneous Item Bias share a common framework. As both procedures are non-parametric, they do not require model calibration (Ackerman, 1992). Both procedures provide tests of significance, are computationally simple, and are inexpensive.

Both procedures typically use the raw score as the conditioning variable to form groups of examinees of comparable ability. For two groups matched on K+1 score categories where K is the number of test items, the MH procedure computes the sum of the odds ratio at each score level weighted by the number of examinees at that score level. Instead of matching on total score, SIB allows the user to select the matching subtest, called the "valid subtest". For examinees who are matched on K "valid" subtest score categories, SIB compares the average proportion correct on the "suspect" subtest for the reference and the focal group examinees. In addition, the SIB procedure, unlike the MH procedure, can simultaneously evaluate DIF present in several test items.

In recent years, the MH procedure has been one of the most popular and widely used procedures for detecting DIF. Research conducted on the Mantel-Haenszel procedure has shown it to be one of the most effective methods for detecting DIF (Hambleton & Rogers, 1989; Raju, Bode & Larsen, 1989; Mazor, Clauser & Hambleton, 1992; Shealy & Stout, 1993). However, recent research has also indicated that under certain circumstances, the MH procedure may have a higher Type I error rate than expected (Zwick, 1990). In general, it appears that the MH procedure has a higher Type I error rate than expected when the probability of a correct response to an item can be described by a two- or a three-parameter item response model rather than a one-parameter model. Roussos and Stout (1993), using simulated data, showed that the observed Type I error rates for the SIB procedure is more acceptable than those of the MH procedure in many realistic 2PL and 3PL IRT models. However, Shealy and Stout (1993) showed that

while adherence to Type I error rate for some realistic 2PL and 3PL models is problematic for the MH procedure, for many other 2PL and 3PL IRT models, both MH and SIB display robustness against Type I error violations. Ackerman (1992) demonstrated that in the multiple-biased item case, the SIB procedure with its emphasis on the selection of a "valid" subtest for matching the examinees, performed better than the MH procedure with total score used as the matching criterion. It should be emphasized that this may be due to the choice of the matching criterion rather than the choice of the procedure.

Research Objectives

Considerable research has been carried out on the MH procedure. On the other hand, the SIB procedure having been developed only recently, has a considerably more limited research base. Although previous research (see for example, Ackerman, 1992; Roussos & Stout, 1993) suggest that the SIB procedure is as promising as the MH procedure in detecting uniform DIF, extending the study to include a number of conditions that have not been investigated previously will determine if under certain circumstances it is more effective than the MH procedure. Therefore, the focus of this study was to investigate in detail the performance of the SIB procedure under a variety of conditions.

The main purposes of this study were to compare the Type I error rates and the power of the MH and the SIB procedures under a variety of conditions to investigate the conditions under which each procedure is optimal for detecting DIF.

Research Design

This research study was conducted on simulated data sets. Only with simulated data sets is it possible to specify different amounts of DIF in a selected number of items and study the power of the MH and SIB procedures on items that are a priori known to be differentially functioning. Examinee response data were simulated under a variety of conditions, with each data set accommodating prespecified levels of a number of different factors that might have an effect on the DIF detection rates. This study was confined to the investigation of uniform DIF because, the MH procedure is designed to detect uniform DIF only.

One factor of interest concerned the size of the sample for the focal and reference groups. Previous research conducted on the impact of sample size on the power of the MH and the SIB procedures suggest that DIF detection rates for both procedures increase with increase in sample size (Mazor et al., 1992; Narayanan & Swaminathan, 1993; Rogers, 1989; Swaminathan & Rogers, 1990, 1993). It is of interest therefore, to investigate the effect of a wide range of sample sizes on the two procedures.

A second factor of interest was the ability distribution differences between the two groups. Mazor et al. (1992) have studied the effects on the MH procedure, when two groups were sampled from equal and unequal distributions. They recommend that, when groups of differing abilities are to be compared, it is probably advisable to use larger sample sizes than might be used when the ability distributions are equal. Shealy and Stout (1993) showed that both MH and SIB procedures display good adherence to the nominal level of

significance even for large target ability differences for a realistic domain of 2PL and 3PL models.

A third factor of interest was the proportion of items exhibiting DIF. In general, a longer test is likely to produce more reliable scores resulting in more reliable ability estimates. On the other hand, increasing the proportion of items exhibiting DIF will produce ability estimates that will be less reliable. When the ability estimates are less reliable, matching will be less accurate. Therefore, the power of the DIF procedures is likely to decrease.

DIF effect size or the amount of DIF contained in an item is the fourth factor that is likely to have an effect on the DIF detection procedures. As DIF effect size increases, the detection rates of the two procedures is expected to increase as well.

The DIF effect sizes were determined using an IRT framework. Within this framework, DIF is said to exist if the ICCs for the two groups are not the same. Therefore, the difference in area between the ICCs for the two groups can be used as a measure of DIF effect size. If the difference between the ICCs is large, then DIF effect size is expected to be large and vice versa. Swaminathan and Rogers (1990) used the area between the ICCs for the two groups of interest as an operational measure of DIF effect size. In their study, they have investigated area values ranging from 0.2 to 0.8. For the purpose of this study, the area between the ICCs for the two groups was used as an operational measure of DIF.

Method

Description of the Power Study

In this study five factors were manipulated: sample size, proportion of items containing DIF, ability distribution differences, DIF effect size and type of item. The three reference group sample sizes (300, 500, 1000) were crossed with the three focal group sample sizes (100, 200, 300) to produce nine sample sizes. Test length was not manipulated but set at 40 items. Standardized achievement and ability tests normally range from about 35 items to about 80 items. The study was confined to a single test length of 40 to investigate the capability of the two procedures to detect DIF in a "short test".

The impact of the differences in underlying ability distributions was investigated by examining three different conditions. These conditions have also been studied by Shealy and Stout (1993). In the first case, the mean of the ability distributions for the two groups was set equal to 0.0 and the standard deviation was set equal to one. This will be referred to as equal ability distribution. In the second case, the mean was set equal to 0.0 and -0.5 for the reference and focal groups respectively, with both standard deviations set equal to one. This will be referred to as unequal(1) ability distribution. Ability distribution that differed by 0.5 standard deviation was specified to simulate the case where there is not a very substantial between group difference. In the third case, the mean was set equal to 0.0 and -1.0 for the reference and the focal groups, respectively, with both standard deviations set equal to one. This will be referred to as unequal(2) ability distribution. Ability distribution that differ by one

standard deviation was chosen to simulate the case where there is a substantial between group difference.

To study the effect of the proportion of items exhibiting DIF, tests were simulated with either 10% or 20% of the items showing DIF. It is seen in practice that standardized achievement tests usually show up to about 10% to 15% items as DIF. The 20% proportion of DIF items was included to represent the "worst case scenario".

For the purpose of this study, the area between the ICCs for the two groups was used to quantify the size of DIF. The areas between the ICCs were computed using the formula given by Raju (1988). Four levels of DIF effect size were chosen equal to the area values .4, .6, .8 and 1.0 to reflect DIF effect sizes ranging from a small amount of DIF to a fairly large amount of DIF. Uniform DIF was simulated by keeping the a-parameters for the two groups the same, but varying the b-parameters for the two groups. This study investigated 24 items showing uniform DIF obtained by varying the level of the common discrimination parameter (low, medium, high), the level of the difficulty parameters for the two groups (low, medium, high) and DIF effect size (area values of .4, .6, .8 and 1.0). In all, six types of item were studied: (1) low b, medium a; (2) low b, high a; (3) medium b, low a; (4) medium b, high a; (5) high b, low a; and (6) high b, medium a. The c-parameters for the 24 DIF items were set equal to .20.

To simulate a test with 10% of the items showing DIF (i.e., four items), and to accommodate the characteristics of items that may affect DIF detection, it was necessary to distribute the 24 DIF items into six 40-item test. Similarly, in order to simulate 20% of the

items showing DIF (i.e., eight items), the 24 DIF items were distributed into three 40-item tests. The non-DIF items were kept the same in all the tests. Item parameter values for the non-DIF items were randomly chosen from published item parameter values from an administration of the Graduate Management and Admissions Test (Kingston, Leary & Wightman, 1988). The c-parameters for all the items were set equal to .20.

Data were generated according to the three-parameter logistic model using the program DATAGEN (Hambleton & Rovinelli, 1973) for a number of tests described above to investigate the capability of the SIB and MH procedures to identify the 24 uniform DIF items described above. The DIF statistics values for the MH procedure were obtained by using the program MHBIAS written by H. Jane Rogers. The SIB DIF statistic values were obtained by using the program SIBTEST written by Shealy, Stout and Roussos. Tables 3.1 and 3.2 present the item parameters values of the DIF and non-DIF items used in the study.

In summary, DIF analyses were carried out for datasets simulated for nine combinations of sample size, three levels of ability distribution differences, two levels of proportion of DIF items, four levels of DIF effect size, and six types of item. In all 1296 conditions were studied. The data were replicated 100 times for each condition.

In computing the MH and SIB DIF statistics, a two-stage procedure recommended by Holland and Thayer (1988) was adopted. In the first-stage, the total score based on all the items was used as the matching criterion to group the examinees and items showing DIF were identified using the MH and the SIB procedures. In the second

stage, items showing DIF (with the exception of the studied item for the MH procedure) were excluded from the calculation of total score used to group examinees. Then the MH and SIB analyses were repeated. The power and Type I error rates (percent of non-DIF items falsely identified as DIF) of the MH and SIB statistics were evaluated at .05 and .01 level of statistical significance.

Results

The Power Study

The DIF detection rates of the MH and SIB procedures as revealed in Tables 3.3 through 3.6 are summarized and presented below.

An analysis of variance (ANOVA) was performed to determine the effects of the five conditions on the performance of SIB and MH procedures. The dependent variable was the number of times the items were identified as DIF in 100 replications of the data. The independent variables were the five different conditions that were manipulated in the study. Table 3.3 shows the ANOVA results for the detection rates across all conditions for the SIB and MH statistics.

A review of ANOVA results shows that for both SIB and MH procedures, sample size, proportion of items containing DIF, type of item and DIF effect size have significant main effects at .05 level of statistical significance. For the SIB procedure, the ability distribution did not have a significant main effect.

Several interaction effects were observed for both procedures. These were sample size with ability distribution differences, sample size with type of item, sample size with DIF effect size, ability distribution differences with type of item, ability distribution

differences with DIF effect size, and type of item with DIF effect size were all significant. For both procedures, there was no interaction effect between percent of DIF and other factors.

Table 3.3 through 3.6 present the mean percent of items correctly identified as differentially functioning for equal, unequal(1) and unequal(2) ability distributions for all conditions. The main findings are as follows:

Effect of Sample Size. For equal, unequal(1) and unequal(2) ability distributions (Tables 3.4 through 3.6), the detection rates for the two procedures showed a steady increase as the sample size increased. In most cases, the SIB procedure identified a slightly higher percentage of DIF items than the MH procedure for unequal ability distributions.

Effect of Type of Item.

1. For equal ability distribution (Table 3.4), the detection rates for the two procedures were highest for highly discriminating/moderate difficulty items followed by highly discriminating/low difficulty items. The lowest detection rates were obtained for high difficulty/low discrimination items followed by high difficulty/medium discrimination items. In general, as the difficulty level of the items increased, the power of the two DIF procedures decreased. On the other hand, as the discrimination level of the items increased, the power of the two DIF procedures increased.
2. The results for unequal(1) and unequal(2) ability distributions (Table 3.5 and 3.6), reveal that for highly discriminating/medium difficulty items, the detection rates for

the two procedures were comparable with those obtained with equal ability distributions. For low difficulty items, the detection rates for both procedures were better than those obtained with equal ability distributions irrespective of the level of discrimination. The detection rates for high difficulty items were lower for both procedures than those obtained with equal ability distribution irrespective of the level of discrimination.

3. A comparison of the detection rates of the two procedures showed that for medium difficulty/low discrimination items, MH identified about 5% to 9% fewer items for unequal(1) and unequal(2) distributions respectively.

4. The detection rates for high difficulty/low discrimination items reduced by about 7% and 15% for unequal(1) and 8% to 30% for unequal(2) distributions respectively for the SIB and MH procedures.

5. For items of high difficulty/medium discrimination, the detection rates for the SIB and MH procedures reduced by 10% and 22% and by 22% and 45% for both unequal(1) and unequal(2) distributions respectively.

6. Overall, the SIB procedure was able to identify more items as DIF for unequal ability distributions than the MH procedure. In fact for certain item types, SIB was able to detect about 25% more items as DIF than MH when the ability distributions were unequal.

Effect of DIF Effect Size. For equal as well as unequal(1) and unequal(2) ability distributions (Tables 3.4 through 3.6), the detection rates for the two procedures steadily increased for increase in the area values from .4 to 1.0 for all sample sizes.

Effect of Proportion of Items Containing DIF. There was an overall decrease of about 1% to 5% for the two procedures as the proportion of items showing DIF increased from 10% to 20%. In general, the detection rates for both procedures showed a similar pattern irrespective of whether tests showed 10% or 20% items as DIF.

The Type I Error Rates

The next step in the analyses was to determine the Type I error rates (number of non-DIF items falsely identified as DIF) for the two procedures. Tables 3.7 through 3.9 present the mean Type I error rates for equal, unequal(1) and unequal(2) ability distributions. The main findings of these tables are:

Effect of Sample Size.

1. Sample size did not seem to affect Type I error rates for both procedures. On the whole, the SIB procedure had a slightly higher Type I error rates than the MH procedure.
2. For equal and unequal(1) ability distributions (Table 3.7 and 3.8), at .05 and .01 levels of significance, the Type I error rates for the MH procedure were the same as the nominal level for all sample sizes. The Type I error rates obtained for the SIB procedure were overall, slightly higher than the nominal level. For unequal(2) ability distribution (Table 3.9), the Type I error rates were inflated for both procedures, the

inflation being slightly higher for the SIB procedure than that of the MH procedure.

Effect of Type of Item. The type of item did not seem to affect the Type I error rates for both procedures. At .05 and .01 levels of significance, the Type I error rates for the MH procedure were the same as the nominal level with a few exceptions. On the whole, the SIB procedure had a slightly higher Type I error rates than those of the MH procedure.

Effect of Proportion of Items Containing DIF.

1. At .05 level of significance, for equal and unequal(1) ability distributions (Table 3.7 and 3.8), the Type I error rates for the MH procedure were within limits for tests with 10% of the items showing DIF and higher than expected in a few cases for tests with 20% of the items showing DIF. For unequal(2) ability distribution (Table 3.9), they were higher than expected in many cases irrespective of whether tests showed 10% or 20% of the items as DIF.
2. For the SIB procedure, the Type I error rates were slightly higher than expected for equal, unequal(1) and unequal(2) ability distributions irrespective of whether test showed 10% or 20% of the items as DIF. The Type I error rates also increased as the ability distribution differences increased and as the proportion of items showing DIF increased.

Discussion

The main findings of the DIF study indicate that, overall, there is high agreement between the SIB and MH procedures in detecting

uniform DIF. As can be expected, the MH and the SIB procedures are affected by sample size. The increase in the power of DIF statistics for increase in sample size is not surprising since the empirical distributions are expected to get closer to the theoretical for increasing sample size. However, the specific purpose of this study was to investigate the effectiveness of these procedures in samples so small that IRT procedures are not feasible. The question therefore becomes, how small a sample size is sufficient for these procedures to be viable methods for detecting uniform DIF.

The results show that detection rates are a function of reference as well as focal group sample sizes for both procedures. Detection rates for the two procedures in this study appear to be more dependent on the focal group sample size than the reference group sample size. In general, on an average, when the focal group sample sizes increased from 100 to 300, the detection rates increased by about 20% whereas, when the reference group sample sizes increased from 300 to 1000, the corresponding increase was only about 10%. These results suggest that varying the sample size and the ratio of reference group to focal group members will have an impact on the performance of MH and SIB procedures for detecting DIF. Overall, a sample size of (300,300) was seen to be sufficient to provide power for the two procedures to detect a reasonable amount of DIF.

These results also suggest that besides sample size, as expected, DIF effect size can have a significant effect on DIF detection procedures irrespective of the size and ratio of reference and focal group members. For all sample sizes, the detection rates both procedures steadily increased as the area values increased from

.4 to 1.0. Overall, there was an increase of only about 10% to 12% in the detection rates for increase in the focal group sample size from 100 to 300 when the area value was 1.0 (high DIF). There was about 26% to 34% increase in the detection rates for increase in the focal group sample size from 100 to 300 when the area value was .4 (low DIF). These numbers were slightly higher for unequal ability distributions. Practitioners should be aware that items which exhibit very small amounts of DIF may go undetected especially when sample sizes are small. However, it can be argued that in such cases, the DIF may be so small that it would make little practical difference.

The results also support the findings of Rogers (1989) that the type of item included is a significant factor influencing the detection rates of the DIF detection procedures. Detection rates were highest for high discrimination items followed by moderate and low discriminating items. Detection rates were lowest for high difficulty items followed by items of moderate difficulty and low difficulty. Highly difficult items will not be answered correctly by the majority of reference and focal group members. Therefore, most difficult items may affect only a small number of examinees since only a very few number of examinees are likely to be found at the extreme ends of the distributions. Fortunately, very difficult items are not very common in standardized achievement tests and hence they may not be a matter of great concern in practice.

The most interesting finding in this study was that the ability distribution differences between the reference and the focal groups did not have an effect on the SIB procedure, whereas, it did have an effect on the MH procedure. The reason for this happening appears to

be due to the regression correction effected in the SIB procedure. According to Shealy and Stout (1993), the regression correction adjusts the studied subtest scores for the two groups so that they are now estimates of the same latent ability in the case of no DIF, even if group target ability distribution differences exist. The SIB procedure can be very useful when differences in the reference and focal group ability distributions exist in practical settings.

The percentage of items exhibiting DIF did not affect the DIF detection rates to a large extent. This may be due to the two-stage procedure adopted in computing the SIB and MH statistics. Items identified as DIF in the first computations were removed when forming the score groups for computing the DIF statistics for the second time. The results of this study suggest the advantage of using the two-stage procedure in computing the DIF statistics, especially when the test contains a large number of items showing DIF. Since the implementation of the two-stage procedure is not difficult in DIF analyses, this method of computing the DIF statistics would be useful to practitioners.

The investigation of the Type I error rates indicate that they were within the nominal limits and conservative for the MH procedure. They were slightly higher for the SIB procedure than those results obtained for the MH procedure for equal ability distributions. There appeared to be inflation of Type I error rates for both procedures as the ability distribution differences increased, the inflation was slightly higher for the SIB procedure. SIB would seem preferable because its Type I error rate is marginally 1% to 2% higher whereas, its power is about 25% higher.

A comparison between Simultaneous Item Bias procedure and the Mantel-Haenszel procedures indicate that the Simultaneous Item Bias procedure is as powerful as the Mantel-Haenszel procedure for detecting uniform DIF when ability distributions are the same and has more power than the Mantel-Haenszel procedure when the reference and focal group ability distributions are unequal. Both procedures are computationally simple, inexpensive and require little computer time. Both methods are therefore interchangeable and can be used under appropriate situations.

Table 3.1

Item Parameters Used to Generate Items with DIF

Item	Item Type	DIF Effect Size	Ref. b1	Foc. b2	Ref. a1	Foc. a2
1	Low b Medium a	.4	-1.80	-1.28	0.90	0.90
2		.6	-1.92	-1.14	0.90	0.90
3		.8	-2.04	-1.01	0.90	0.90
4		1.0	-2.16	-0.88	0.90	0.90
5	Low b High a	.4	-1.80	-1.28	1.25	1.25
6		.6	-1.92	-1.14	1.25	1.25
7		.8	-2.04	-1.01	1.25	1.25
8		1.0	-2.16	-0.88	1.25	1.25
9	Medium b Low a	.4	-0.26	0.26	0.50	0.50
10		.6	-0.39	0.39	0.50	0.50
11		.8	-0.51	0.51	0.50	0.50
12		1.0	-0.64	0.64	0.50	0.50
13	Medium b High a	.4	-0.26	0.26	1.25	1.25
14		.6	-0.39	0.39	1.25	1.25
15		.8	-0.51	0.51	1.25	1.25
16		1.0	-0.64	0.64	1.25	1.25
17	High b Low a	.4	1.28	1.80	0.50	0.50
18		.6	1.14	1.92	0.50	0.50
19		.8	1.01	2.04	0.50	0.50
20		1.0	0.88	2.16	0.50	0.50
21	High b Medium a	.4	1.28	1.80	0.90	0.90
22		.6	1.14	1.92	0.90	0.90
23		.8	1.01	1.24	0.90	0.90
24		1.0	0.88	2.16	0.90	0.90

Table 3.2

Item Parameters for the Non-DIF Items

Item	b	a	c	Item	b	a	c
1	-0.30	.44	.20	19	1.09	.55	.20
2	-1.06	.55	.20	20	1.64	1.40	.20
3	1.02	.82	.20	21	1.13	.92	.20
4	-1.96	.52	.20	22	-1.55	.64	.20
5	1.28	1.02	.20	23	.81	1.01	.20
6	.61	.82	.20	24	-.53	.61	.20
7	.42	.92	.20	25	1.05	.70	.20
8	1.68	.65	.20	26	.64	1.02	.20
9	-2.70	.56	.20	27	2.12	.48	.20
10	-1.39	.29	.20	28	.91	1.01	.20
11	-1.12	.35	.20	29	.87	.53	.20
12	-1.37	.31	.20	30	-2.63	.36	.20
13	.10	1.05	.20	31	-1.21	1.12	.20
14	-.09	.51	.20	32	-.57	.86	.20
15	.61	.73	.20	33	-1.29	.59	.20
16	.95	.88	.20	34	.40	.56	.20
17	-.35	1.11	.20	35	1.11	1.09	.20
18	.57	1.32	.20	36	-.93	.88	.20

Note: Item parameters for items 1-36 did not vary across conditions.

Table 3.3

Analysis of Variance of the Effects of all Factors on the Performance of the Simultaneous Item Bias and Mantel-Haenszel Procedures on DIF

Factor	SIB F	SIB p	MH F	MH p
Main Effects				
Sample Size	273.60	.000*	209.95	.000*
Ability Distribution	0.65	.520	260.50	.000*
Percent DIF	31.95	.000*	39.49	.000*
Type of Item	1737.39	.000*	2878.89	.000*
DIF Effect Size	1958.71	.000*	1857.50	.000*
Interaction Effects				
Sample Size X Ability Distribution	3.32	.000*	2.83	.000*
Sample Size X Percent of DIF	.30	.992	.22	.986
Sample Size X Type of Item	10.27	.000*	6.84	.000*
Sample Size X DIF Effect Size	5.63	.000*	3.03	.000*
Ability Distribution X Percent of DIF	.03	.975	.02	.980
Ability Distribution X Type of Item	76.64	.000*	184.12	.000*
Ability Distribution X DIF Effect Size	42.58	.000*	38.52	.000*
Percent of DIF X Type of Item	.99	.423	1.73	.124
Percent of DIF X DIF Effect Size	1.93	.123	0.69	.560
Type of Item X DIF Effect Size	69.62	.000*	73.73	.000*

Table 3.4

Mean Percent Detection Rates of the Simultaneous Item Bias and Mantel-Haenszel Procedures for Equal Ability Distributions Under all Conditions

Factor			10% DIF				20% DIF			
			SIB		MH		SIB		MH	
		α =	.05	.01	.05	.01	.05	.01	.05	.01
Sample Size			(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
Ref.	Foc.									
300	100		62	45	62	47	60	43	60	44
300	200		78	67	77	64	74	61	72	58
300	300		84	72	82	70	81	70	78	66
500	100		62	46	64	50	61	48	63	48
500	200		81	69	80	69	79	68	77	65
500	300		87	79	87	76	83	72	84	76
1000	100		66	49	69	55	65	48	67	52
1000	200		84	73	85	74	82	70	82	71
1000	300		90	82	90	82	88	78	88	79
Type of Item										
Low b	Medium a		85	71	85	56	81	68	83	72
Low b	High a		88	76	89	81	85	75	86	79
Medium b	Low a		73	59	73	59	70	55	70	54
Medium b	High a		95	90	95	93	93	87	94	88
High b	Low a		58	40	56	36	55	37	51	34
High b	Medium		66	52	64	48	66	50	63	45
DIF Effect Size										
Area										
.4			50	32	49	32	46	27	45	28
.6			75	59	76	61	72	56	72	56
.8			88	79	88	78	87	77	87	76
1.0			95	89	95	90	95	88	94	87

Table 3.5

Mean Percent Detection Rates of the Simultaneous Item Bias and Mantel-Haenszel Procedures for Unequal(1) Ability Distributions Under all Conditions

Factor		10% DIF				20% DIF			
		SIB		MH		SIB		MH	
	α =	.05	.01	.05	.01	.05	.01	.05	.01
Sample Size		(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
Ref.	Foc.								
300	100	61	47	58	45	59	43	56	42
300	200	74	62	70	55	74	60	70	57
300	300	82	71	77	67	79	67	72	61
500	100	64	51	62	49	60	48	60	48
500	200	80	69	75	65	80	68	74	62
500	300	86	77	81	72	85	76	79	68
1000	100	67	54	65	51	62	51	61	50
1000	200	84	74	78	64	82	72	76	65
1000	300	89	81	85	77	88	80	82	74
Type of Item									
Low b	Medium a	91	80	92	83	89	78	91	83
Low b	High a	96	91	94	86	89	82	94	89
Medium b	Low a	73	58	69	55	69	57	67	50
Medium b	High a	95	90	93	90	93	88	93	87
High b	Low a	51	34	41	24	50	33	37	20
High b	Medium a	55	38	42	25	54	37	39	21
DIF Effect Size									
Area									
.4		55	38	50	35	52	34	47	33
.6		75	61	70	57	72	57	67	54
.8		85	75	81	71	85	74	79	68
1.0		92	84	89	81	91	84	87	78

Table 3.6

Mean Percent Detection Rates of the Simultaneous Item Bias and Mantel-Haenszel Procedures for Unequal(2) Ability Distributions Under all Conditions

Factor		10% DIF				20% DIF			
		SIB		MH		SIB		MH	
	α =	.05	.01	.05	.01	.05	.01	.05	.01
Sample Size		(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
Ref.	Foc.								
300	100	66	54	51	45	63	53	52	42
300	200	76	63	63	53	74	65	63	53
300	300	80	69	69	60	78	69	67	58
500	100	68	54	59	48	66	54	56	45
500	200	80	70	67	59	79	69	66	57
500	300	86	78	74	65	84	75	71	62
1000	100	70	58	60	50	68	55	58	48
1000	200	82	73	70	62	80	71	68	59
1000	300	87	79	75	68	87	78	71	64
Type of Item									
Low b	Medium a	97	91	96	91	96	92	95	90
Low b	High a	99	97	99	97	99	95	98	94
Medium b	Low a	77	61	64	46	75	59	59	42
Medium b	High a	95	90	92	85	95	89	90	81
High b	Low a	50	33	26	13	46	31	22	10
High b	Medium a	44	26	19	8	42	27	17	7
DIF Effect Size									
Area									
.4		62	48	50	40	59	46	47	36
.6		74	63	63	54	73	62	61	51
.8		83	73	72	63	81	73	70	61
1.0		89	81	80	72	88	80	77	68

Table 3.7

Mean Percent Type I Error Rates of the Simultaneous Item Bias and Mantel-Haenszel Procedures for Equal Ability Distributions Under all Conditions

Factor			10% DIF				20% DIF			
			SIB		MH		SIB		MH	
		α =	.05	.01	.05	.01	.05	.01	.05	.01
Sample Size			(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
Ref	Foc									
300	100		6.2	1.5	3.7	0.7	6.7	1.6	4.2	0.8
300	200		5.2	1.0	3.6	0.6	6.0	1.4	4.5	0.8
300	300		5.4	1.3	4.2	0.8	5.8	1.3	4.5	1.0
500	100		6.0	1.6	3.6	0.6	7.6	2.4	4.0	0.8
500	200		5.2	1.1	3.8	0.6	6.1	1.4	4.7	0.9
500	300		5.6	1.1	4.2	0.6	6.6	1.6	5.4	1.1
1000	100		6.3	2.0	3.8	0.8	7.7	2.6	4.2	0.8
1000	200		6.0	1.4	4.2	0.8	6.8	1.9	4.5	0.9
1000	300		5.5	1.1	4.2	0.8	6.4	1.5	5.1	1.0
Type of Item										
Low b	Medium a		5.7	1.9	3.9	0.8	9.1	3.0	5.3	1.1
Low b	High a		6.6	1.9	3.0	0.7	6.3	3.1	4.4	1.0
Medium b	Low a		6.2	1.7	3.6	0.7	6.3	1.7	3.7	0.5
Medium b	High a		5.3	1.2	4.2	0.8	6.1	1.3	5.2	0.8
High b	Low a		4.8	1.0	3.6	0.6	5.6	1.2	4.3	0.7
High b	Medium a		6.0	1.1	4.2	0.7	6.9	1.4	4.5	1.1

Table 3.8

Mean Percent Type I Error Rates of the Simultaneous Item Bias and Mantel-Haenszel Procedures for Unequal(1) Ability Distributions Under all Conditions

Factor			10% DIF				20% DIF			
			SIB		MH		SIB		MH	
		α =	.05	.01	.05	.01	.05	.01	.05	.01
Sample Size			(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
Ref	Foc									
300	100		6.1	1.6	3.6	0.6	6.6	1.9	4.1	0.8
300	200		5.8	1.3	3.8	0.7	5.8	1.4	4.8	0.9
300	300		5.7	1.3	4.2	0.9	5.7	1.3	4.7	0.9
500	100		6.5	1.2	3.9	0.8	6.9	2.3	4.6	0.9
500	200		5.5	1.6	3.8	0.7	6.2	1.6	4.9	1.0
500	300		5.8	1.2	4.5	0.9	6.5	1.7	7.7	3.7
1000	100		6.7	1.9	4.2	0.9	7.4	2.4	4.2	0.8
1000	200		5.8	1.5	4.2	0.9	6.0	1.5	4.9	1.0
1000	300		5.7	1.3	4.4	0.8	6.0	1.4	5.2	1.4
Type of Item										
Low b	Medium a		6.1	1.4	4.0	0.8	5.9	1.6	4.1	1.2
Low b	High a		5.8	1.4	5.1	0.8	5.9	1.7	5.2	1.0
Medium b	Low a		6.0	1.7	4.4	0.8	5.8	1.6	5.4	1.7
Medium b	High a		6.7	1.3	4.0	0.5	7.1	1.9	6.1	1.5
High b	Low a		5.8	1.4	4.8	1.1	6.9	1.7	6.0	1.4
High b	Medium a		5.5	1.4	4.6	0.9	5.9	1.4	4.7	1.4

Table 3.9

Mean Percent Type I Error Rates of the Simultaneous Item Bias and Mantel-Haenszel Procedures for Unequal(2) Ability Distributions Under all Conditions

Factor			10% DIF				20% DIF			
			SIB		MH		SIB		MH	
		α =	.05	.01	.05	.01	.05	.01	.05	.01
Sample Size			(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
Ref	Foc									
300	100		6.8	2.0	4.2	0.9	7.8	2.2	4.1	1.0
300	200		7.2	2.0	4.8	1.0	8.6	2.5	4.7	1.1
300	300		9.1	2.7	4.9	1.0	10.0	3.2	5.5	1.2
500	100		7.2	2.1	4.2	0.9	7.8	2.3	5.1	1.1
500	200		8.2	2.1	5.0	1.1	9.0	2.5	5.6	1.1
500	300		9.3	2.8	5.5	1.2	10.2	3.2	6.1	1.3
1000	100		7.8	2.3	4.5	1.0	8.0	2.6	4.8	1.0
1000	200		8.0	2.1	5.7	1.4	8.7	2.5	6.0	1.4
1000	300		9.4	2.5	6.2	1.6	10.2	3.2	7.2	2.1
Type of Item										
Low b	Medium a		6.8	3.1	5.6	1.3	7.8	3.1	6.1	1.0
Low b	High a		7.2	2.3	5.9	1.3	8.0	2.5	6.3	3.5
Medium b	Low a		7.0	2.1	3.1	0.6	6.7	1.3	3.4	1.0
Medium b	High a		6.7	2.0	5.0	1.1	7.3	2.0	5.9	1.5
High b	Low a		6.2	2.2	5.9	1.3	6.7	1.9	7.4	2.0
High b	Medium a		7.1	2.4	5.6	1.1	6.7	1.9	6.4	1.7

C H A P T E R IV

THE DISTRIBUTIONAL PROPERTIES OF THE MANTEL-HAENSZEL AND THE SIMULTANEOUS ITEM BIAS DIF STATISTICS

Introduction

Statistical procedures for detecting DIF investigate whether there are observed differences between the performances of different groups defined by ethnic background, culture and gender. Although a variety of statistical procedures for detecting DIF are available, the focus of current research in DIF is in the development of procedures that are simple, cost-effective, and easy to implement in practice.

This study investigated two currently popular non-parametric DIF detection procedures, the MH and the SIB. These two procedures share a common framework. In both procedures, DIF can be studied within the item response theory (IRT) framework. From an IRT perspective, DIF is said to exist when examinees of the same ability but belonging to different groups have different probabilities of answering an item correctly. Although both procedures have IRT justifications, both procedures are attractive because they are easy to implement in practice. Both procedures are computationally simple, inexpensive and provide statistics that have associated tests of significance. Under the null hypothesis, the MH statistic has a chi-square distribution with one degree of freedom that is modified with a continuity correction to improve the accuracy of the chi-square percentage points as approximation to the observed significance levels (Holland & Thayer, 1988). The SIB statistic has a normal distribution with mean zero and standard deviation equal to one under the null hypothesis.

The SIB procedure has an added advantage: it is capable of detecting DIF in one or more test items.

A number of studies have compared the power, Type I error rates, and the distributional properties of the MH procedure with other popular DIF detection procedures (Rogers, 1989; Swaminathan & Rogers, 1990; Narayanan & Swaminathan, 1993). These studies have indicated that the MH procedure performed well in detecting uniform DIF and produced lower Type I error rates than the other procedures. However, investigations of the distributional assumptions of the MH statistic have shown that its distributional properties are not as readily met as those of the LR and the SIB procedures. Rogers (1989) examined the distributional assumptions of the MH statistic under 10 conditions (100 replications each) for data generated using a three-parameter logistic model and showed that they were not satisfied for many conditions. However, Narayanan and Swaminathan (1993) showed that the MH statistic was not distributed at all as a chi-square for any of the 45 conditions (1000 replications each) examined. These results raise questions about the practice of using MH (with the continuity correction) as a test statistic for detecting DIF. Therefore, one of the main purposes of this research was to determine the impact of the continuity correction on the distributional assumptions and the power of the MH statistic.

Research Objectives

While a number of research studies have investigated the power and Type I error rates of the MH and the SIB procedures, relatively little research has been conducted on the investigation of the

distributional assumptions of the MH statistic (Rogers, 1989; Rogers & Swaminathan, 1993), and no research whatsoever has been done on the SIB statistic on this aspect. The investigation of the distribution of a test statistic is necessary because, it will indicate to what extent the distributional assumptions of the statistic are satisfied. If they are violated, then the interpretations based on the results from the use of the statistic for DIF studies are not valid.

The main purposes of the study were (a) to investigate the conditions under which the asymptotic distributions of the Mantel-Haenszel statistic (with and without the continuity correction) and the SIB statistic were obtained, and (b) to determine the Type I error rates and power of Mantel-Haenszel and the SIB statistics.

Method

The study was conducted in two parts. Part one investigated whether or not the distributional properties of the SIB and MH test statistics are correct. Therefore, the research questions were (1) to determine whether the MH statistic was distributed as a chi-square distribution with one degree of freedom and (2) to determine whether the SIB statistic was distributed normally with mean zero and standard deviation one. If the expected distributions of the MH and SIB statistics were not obtained, then the validity of the test statistics as indicators of the presence of DIF would be in question.

Studies investigating the distributional assumptions of the MH statistics have shown that, in a large number of studied conditions, the statistic was not distributed as expected. Furthermore, in most conditions, the estimated means and standard deviations of the

empirical sampling distributions were below their expected values (mean = 1.0, S.D. = 1.414). It is possible that the means and the standard deviations of the MH statistic are being underestimated as the result of the inclusion of the continuity correction. Therefore, to determine the impact of the continuity correction on the MH chi-square statistic, this study examined the asymptotic distributional assumptions of two variations of the MH statistic, namely, MH(1) (with the continuity correction) and MH(2) (without the continuity correction).

Part two of the study investigated the power of SIB, MH(1) and MH(2) statistics to determine their potential for detecting uniform DIF in test items. The Type 1 error rates (false positive rates) of the SIB, MH(1) and MH(2) statistics for the items showing non-DIF were determined and compared.

This research study was conducted on simulated data sets. Data sets with simulated examinee responses have the advantage of being able to accommodate a number of factors that can have an impact on the distributional assumptions and power of the statistic of interest so that investigations of the distributional assumptions and power can be carried out under the desired conditions.

Description of the Distribution Study

The investigation of the distributional assumptions of the SIB, MH(1) and MH(2) statistics were carried out by manipulating several factors that can affect the distributional properties. Since the distributional properties are asymptotic, it was expected that as the sample size increased, the empirical sampling distribution of a test

statistic would more likely approach the theoretical distribution. Therefore, sample sizes were manipulated to study their effect on the asymptotic distributional properties of the three statistics.

In practice, it is seen that examinees in the focal groups may be a small number often ranging from about 100 to 300 examinees. Therefore, three levels of reference group sample sizes (300, 500, 1000) were crossed with three levels of focal group sample sizes (100, 200, 300) to give a total of nine sample sizes. Ability values for the two groups were randomly sampled from a normal distribution with mean zero and standard deviation equal to one.

Data for the study were simulated for a test length of 45 items (which is approximately the average length of the subtests of many standardized tests). Nine out of the 45 items were studied for the distributional properties of the three test statistics. The item parameters values for these nine items were obtained by crossing three levels of difficulty parameters (low (-1.5), medium (0.0), high (1.5)) with three levels of discrimination parameters (low, (0.5), medium (1.0), high (1.5). The c-parameters for the nine items were set equal to .20. In all, a total of 81 conditions, obtained by crossing nine levels of sample size with nine types of item, were studied for the distributional assumptions of the SIB and MH(1) and MH(2) test statistics.

For the remaining set of 36 items in the 45-item test, to represent a realistic situation, item parameter values were set to values randomly chosen from published tables of item parameter values obtained from an administration of the Graduate Management Admission Test (Kingston, Leary & Wightman, 1988). Data for the study were

simulated using the three-parameter logistic model using the program DATAGEN (Hambleton & Rovinelli, 1973). For each of the nine sample sizes, item response data for 45 items described above were simulated one each for the reference and the focal groups. The same item parameter values were specified for the reference and the focal groups to represent items in which no DIF were present.

The distributions of the test statistics for the SIB, MH(1) and MH(2) statistics across 1000 replications of the data were obtained. To test the asymptotic properties of the SIB, MH(1) and MH(2) statistics, the Kolmogorov-Smirnov (K-S) and the Wilks-Shapiro (W-S) tests were carried out wherever appropriate. The K-S goodness-of-fit test would indicate if the MH(1) and MH(2) statistic have chi-square distributions with one degree of freedom and if the SIB statistic has a normal distribution with a mean zero and standard deviation one. The Wilks-Shapiro goodness-of-fit test would also indicate if the conditions for the normality of a distribution is satisfied and is therefore, appropriate for the SIB statistic.

The Kolmogorov-Smirnov Test. The Kolmogorov-Smirnov test is similar to a chi square goodness-of-fit test in that it tests for significant differences between observed and an expected frequency distribution. In this test, the observations are first ordered and standardized and their observed cumulative frequencies are computed. Their expected cumulative distributions are calculated under the assumption that the theoretical distribution is satisfied. The maximum absolute differences between the observed and the expected cumulative frequencies are computed. If this value exceeds the

tabulated value, the null hypothesis that the sampling distribution has the same form as the theoretical distribution is rejected.

The Wilks-Shapiro Test. The Wilks-Shapiro (W-S) test provides a general test to determine if the distributions to be tested are normal (bell-shaped). In a normal probability plot, the W-S test tests the adequacy of the linear fit. To obtain the W-S statistic, the sample values to be tested for normality of assumptions are first ordered. Then a regression of the ordered sample values on the corresponding expected normal order statistics is carried out. For a sample of normally distributed values, a linear fit is expected if the null hypothesis is true. The W-S statistic value is obtained as an F-ratio from the generalized least-square analysis to judge the adequacy of the linear fit.

From the test statistic values obtained from the distribution results, the number of false positives (non-DIF items incorrectly identified as DIF) were determined for the SIB, MH(1) and MH(2) statistics at the 95th and 99th percentile cutoffs. The number of false positive errors would indicate whether or not the nominal Type I error rates were obtained.

Description of the Power Study

In Chapter III, the power of the MH and the SIB statistics was investigated under a variety of conditions. The power study was repeated in this investigation to determine the impact of the continuity correction on the MH statistic in detecting DIF and also to compare the results with the results of the SIB statistic.

In this phase, the power of the SIB, MH(1) and MH(2) statistics was studied with data simulated to reflect a variety of conditions likely to have an impact on the detection of uniform DIF in test items. Uniform DIF was simulated by choosing different values for the difficulty parameters for the reference and the focal groups while keeping the discrimination and the lower asymptote parameters for the two groups the same. The power of the SIB, MH(1) and MH(2) statistics were examined on items that were a priori known to be differentially functioning.

Rogers (1989) identified a number of factors that might have an impact on the DIF detection rates of the DIF procedures. In this study, six such factors were manipulated: sample size, test length, proportion of items containing DIF, ability distribution differences, DIF effect size and type of item were manipulated. The three reference group sample sizes (300, 500, 1000) were crossed with the three focal group sample sizes (100, 200, 300) to produce nine sample sizes. Since standardized achievement and ability tests usually range from about 35 items to 80 items, two test lengths, a 40-item test to represent a "short test" and a 60-item test to represent a "medium test' were simulated.

To investigate the impact of ability distribution differences on the detection rates, two conditions were simulated. In the first condition, the ability distributions for the two groups were set to be equal with mean 0.0 and standard deviation equal to one to allow for comparisons of examinees of equal ability. In the second condition, the mean of the reference and the focal groups were set to be equal to 0.0 and -1.0 respectively and the standard deviations for both groups

were set equal to one. Unequal ability distributions in which the reference and the focal group examinees differ by one standard deviation were chosen to simulate the condition commonly found in practice in many testing situations. To study the impact of the proportion of items containing DIF, tests were simulated with either 10% or 20% of the items containing DIF. Although in practice standardized achievement test usually contain up to about 10% items as DIF, the 20% case was included to represent the "worst case scenario".

Four levels of DIF effect size were chosen equal to the area values of .4, .6, .8 and 1.0 using the formula given by Raju (1988). Uniform DIF was simulated by keeping the a-parameters the same for the two groups, but varying the b-parameters for the two groups. In this study, six types of items were obtained by varying the level of the common discrimination parameter (low, medium, high), the level of the difficulty parameters for the two groups (low, medium, high). The six types of items that were chosen were: (1) low b, medium a; (2) low b, high a; (3) medium b, low a; (4) medium b, high a: (5) high b, low a; (6) high b, medium a. Twenty four items were obtained by crossing four levels of DIF effect size (area values of .4, .6, .8, and 1.0) with the six types of items described above.

To simulate a 40-item test with 10% of the items showing DIF (i.e., four items) and to accommodate the characteristics of items that may affect DIF detection, it was necessary to distribute the 24 items into six 40-item tests. Similarly, in order to simulate 20% of the items showing DIF (i.e., eight items), the 24 DIF items were distributed into three 40-item tests.

To simulate a 60-item test with 10% of the items showing DIF (i.e., six items), it was necessary to distribute the 24 items into four 60-item tests. Similarly, in order to simulate 20% of the items showing DIF (i.e., eight items), the 24 DIF items were distributed into three 40-item tests. Item parameter values for the non-DIF items were kept the same in all the 40-item and 60-item tests. They were values randomly chosen from published item parameter values obtained from an administration of the Graduate Management and Admissions Test (Kingston, Leary & Wightman, 1988). The c-parameters for all the items were set equal to .20. Table 4.1 presents the item parameter values specified for the distribution and the DIF studies.

Data were generated using the program DATAGEN (Hambleton & Rovinelli, 1973) for a number of tests described above to investigate the capability of the SIB and MH procedures to identify items that are a priori known to be differentially functioning.

In summary, DIF analyses were carried out with data sets simulated for nine levels of sample size, two levels each of test length, ability distribution differences, proportion of items containing DIF, four levels of DIF effect size, and six type of items. In all, 1728 conditions were studied. The data were replicated 100 times for each condition. The power and Type 1 error rates of the three statistics were evaluated at .05 and .01 levels of significance.

Results

The Distribution Study

The Kolmogorov-Smirnov (K-S) and Wilks-Shapiro (W-S) test results for investigating the distributional properties of the SIB and

MH statistics are presented in Tables 4.2 through 4.4, respectively. The critical value of the K-S test statistic at the .05 level of significance for 1000 simulations was .043. The W-S critical value at .05 level of statistical significance was 0.983. In other words, a statistic is significant if the probability level is less than .05. The main findings are as follows:

The Distribution of the SIB Statistic.

1. The results of the SIB statistic are presented in Table 4.2. Table 4.2 reveals that the estimated means and the standard deviations of most of the 81 empirical distributions closely approximated the mean (0.0) and the standard deviation (1.0) of the theoretical distributions (normal).
2. The K-S goodness-of-fit results show that the theoretical distributions were obtained for all but 10 of 81 conditions. Eight of these occurrences were for focal group sample sizes of 100 and one each for focal group sample size of 200 and 300. Six of the 10 conditions occurred for items with high difficulty and two each for items with low and medium difficulty.
3. The W-S goodness-of-fit results (Table 4.2) confirm the K-S test results, and provide further evidence that the normality assumptions of the SIB test statistic were satisfied for all but seven of 81 conditions. Five of these occurrences were obtained for focal group sample size of 100 and two for a focal group sample size of 200. Five of the occurrences were common to both tests. The W-S test does not specifically test for conditions of normality with mean zero and standard deviation 1.0.

The Distribution of the MH Statistic.

1. The results of the MH statistic are presented in Table 4.3. Table 4.3 shows that the estimated means and the standard deviations of the empirical chi-square distributions of the MH(1) statistic for most of the 81 items were lower than the mean (1.0) and the standard deviation (1.414) of the theoretical chi-square distribution. The estimated means and the standard deviations of the MH(2) statistics closely approximated those of the theoretical chi-square distribution.
2. The MH(1) and MH(2) statistics were not distributed as a chi-square distribution under any of the 81 studied conditions. However, the empirical distributions of MH(2) more closely approximated their expected distributions than MH(1) for all the studied conditions.

The Type I Error Rates of the SIB and the MH Statistics

1. The observed Type I error rates of the SIB statistic were higher than the expected limits (Table 4.4). At the .05 level of significance, the Type I errors rates for the SIB statistic ranged from about 4.4% to about 7.8%. At the .01 level of significance, they ranged from about 0.4% to about 2.5%. The Type I error rates for MH(1) statistic were well within acceptable limits and quite conservative. At the .05 level of significance, the Type I error rates varied from 2.2% to about 5.0% and at the .01 level of significance, they were between 0.1% to 1.2%. The Type I error rates for the MH(2) statistic were higher than those of MH(1), but on the whole, most of them

were within acceptable limits. At .05 level of significance, they ranged from about 3.5% to about 6.5% and at .01 level of significance, they ranged from 0.5% to 1.5%.

2. Although the Type I error rates of the SIB and MH(2) statistics were higher than those of the MH(1) statistic, comparatively, MH(2) showed better results than SIB.

The Power Study

Table 4.5 presents the mean detection rates for the three test statistics under all conditions, namely, sample size, test length, ability distribution difference, proportion of items containing DIF, type of item and DIF effect size. The results evaluated at the .05 level of significance are summarized and reported in the following sections. The main findings in Table 4.5 are as follows:

Effect of Sample Size. The detection rates for the SIB, MH(1) and MH(2) statistics showed a steady increase for increases in the three levels of the focal group sample sizes.. For a focal group sample size of 100 examinees, they were about 64% (SIB), 60% (MH(1)), and 63% (MH(2)), but increased to about 83% (SIB), 78% (MH(1)), and 81% (MH(2)) for a focal group sample size of 300. The highest detection rates were seen for the SIB statistic followed by the MH(2) and MH(1) statistics. For all three statistics, the mean percent detection rates for a reference group sample size of 300 as well as 1000 examinees were about the same. The number of examinees in the reference group did not seem to have an impact on DIF detection rates.

Effect of Test Length. Test length did not appear to have any marked effect on the detection rates of the SIB, MH(1) and MH(2) statistics. The detection rates of the SIB, MH(1) and MH(2) statistics were about 75%, 70% and 73% respectively for test lengths of 40 and 60 items. Again, the MH(2) statistic was able to identify about 3% more items as DIF than MH(1) statistic with SIB having the highest detection rates.

Effect of Ability Distribution Difference. All three statistics were able to identify more items as DIF for equal ability distribution than unequal ability distribution. For equal ability distribution, the detection rates for the SIB, MH(1) and MH(2) statistics were 77%, 76% and 77%, respectively. These numbers for unequal ability distribution were 74%, 64% and 69%, respectively, for the three statistics. The SIB statistics was able to identify about 5% to 10% more items as DIF than the other two statistics for unequal ability distribution.

Effect of Proportion of Items Containing DIF. There was a small decrease in the detection rates of the three statistics as the proportion of DIF items in the test increased from 10% to 20%. For tests with 10% of the items showing DIF, the detection rates were 75%, 71% and 73% respectively for the three statistics. These numbers for tests with 20% of the items showing DIF were 74%, 68%, and 70% respectively.

Effect of DIF Effect Size. The detection rates of the SIB, MH(1) and MH(2) statistics showed a steady increase for increase in the area values from .4 to 1.0. This increase ranged from 52% to 91% for the SIB statistic, 47% to 86% for MH(1) statistic and 50% to 89%

for MH(2) statistic. There was a significant increase of about 40% in the detection rates for the three statistics when the area values increased from .4 to 1.0.

Effect of Type of Item. The detection rates for the three statistics were highest for highly discriminating/moderate difficulty items followed by highly discriminating/low difficulty items. The lowest detection rates were for high difficulty/low discrimination items followed high difficulty/medium discrimination items. In general, as the difficulty level of the items increased, the power of the two DIF procedures decreased. On the other hand, as the discrimination level of the items increased, the power of the two DIF procedures increased.

The Type I Error Rates of the SIB and the MH Statistics

The results of the investigation of the Type I error rates (number of non-DIF items falsely identified as DIF) for the three statistics are presented in Table 4.6. The main findings are:

Effect of Sample Size.

1. Sample size did not seem to affect the Type I error rates for all three statistics. On the whole, the SIB procedure had the highest Type I error rates followed by MH(2) and MH(1) statistics.
2. The Type I error rates for the MH(1) statistics were within acceptable limits for all sample sizes. For the MH(2) statistics, they were slightly inflated (about 5.5% to 6.1%). The inflation was highest for the SIB statistic ranging from about 6.5% to 7%.

Effect of Test Length. Test length did not seem to affect the Type I error rates of the three statistics. The Type I error rates were within limits for MH(1), followed by MH(2) (about 5.6%) and SIB (about 7%) statistics.

Effect of Ability Distribution Difference. The Type 1 error rates for equal and unequal ability distribution were within limits for MH(1) and about 5.5% for MH(2), whereas, for the SIB statistics, they were higher for unequal ability distribution (about 7%) than for equal ability distribution (about 6%).

To investigate the impact of the continuity correction on the MH statistic, the detection rates and the nominal Type 1 error rates of the MH(1) and MH(2) statistics averaged over 100 replications were determined for two types of item with high difficulty for the nine sample sizes. The reason for choosing high difficulty items was the low detection rates observed with such items in the previous analyses. The results of these analyses are presented in Tables 4.7 through 4.10.

Table 4.7 and 4.8 reveal the detection rates of MH(1) and MH(2) statistics for two types of high difficulty items and different sample sizes in the 40-item tests containing 10% items as DIF for equal and unequal ability distributions. For both types of items, the MH(2) statistic was able to identify more items as DIF than MH(1) statistic for all sample sizes. Overall, there was an increase of about 1% to about 5% over the detection rates of the MH(1) statistic.

Table 4.9 and 4.10 present the Type I error rates of the 36 non-DIF items in the 40-item with 10% of the items showing DIF for MH(1) and MH(2) statistics. The Type I error rates of the MH(2) statistic

were on the whole higher than those of MH(1) statistic. The Type I error rates of the MH(1) statistic ranged from about 3.5% to about 4.5% and about 3.8% to about 6.7% for equal and unequal ability (2) distributions respectively at the .05 level of significance. For the MH(2) statistic, these values ranged from about 4.5% to about 5.7% and 5.4% to about 7.7%, respectively.

Discussion

The results of the first part of the study indicate that for most types of items, the SIB statistic has the expected distribution (normal with mean zero and standard deviation equal to one) for the reference and focal group for all sample sizes. Items for which the theoretical distributions were not obtained were mainly highly difficult and/or highly discriminating items. The MH(1) and MH(2) statistics did not appear to be distributed as a chi-square distribution with one degree freedom for all the studied conditions. However, the results of the MH(2) statistics were closer to the critical value of the K-S test statistic than those of the MH(1) statistic.

The results also suggest that the estimated means and standard deviations of the distributions of the SIB and MH(2) statistics are more acceptable than those of the MH(1) statistic values. For most conditions, the means and standard deviations of the MH(1) statistic are lower than the expected values (mean = 1.0, S.D. = 1.414) for most conditions. Investigation of the Type I error rates of the MH(1) and MH(2) show that they are within the nominal levels, whereas, they are about 1% to 3% higher than the nominal limits for the SIB statistic.

Summarizing the asymptotic results, the SIB statistic appears to be conforming quite well to the asymptotic theory for all the studied conditions. Although its Type I error rate is marginally higher than the nominal limits, the use of the SIB statistic for detecting DIF in the practical settings is recommended because it satisfies the distributional assumptions on which it is based. Although both MH(1) and MH(2) statistics do not seem to have their expected distributions, their Type 1 error rates were within acceptable ranges. Because the MH(2) statistic tends to conform more readily to the underlying theory and has better estimated means and standard deviations than MH(1), it is recommended that it be used in preference to the MH(1) statistic.

The results of the power study suggest that the SIB statistic and the two variations of the MH statistic are about equally effective in detecting uniform DIF. However, there was high agreement between the detection rates of the SIB and MH(2) statistic. It is obvious that the removal of the continuity correction in the computation of the MH statistic increased the MH(2) chi-square test statistic values. With higher estimated chi-square values, the MH(2) statistic was able to identify more items as DIF than the MH(1) statistic.

As revealed in previous research (Rogers, 1989; Mazor et el., 1992), all three statistics showed a marked increase in the detection rates for increase in sample size. On an average, when the focal group sample sizes increased from 100 to 300, the detection rates increased by about 20% whereas, when the reference group sample sizes increased from 300 to 1000, the corresponding increase was minimal. Overall, a sample size of 300 in the focal group was seen to be sufficient to provide power for the two procedures to detect a

reasonable amount of DIF, irrespective of the size of the reference group sample size. In practical settings where the focal group sample size is likely to be small, these three statistics will be very useful for DIF detection purposes.

As expected, the size of DIF also seem to have a significant effect on DIF detection rates irrespective of the other factors. For all sample sizes, the detection rates for both procedures steadily increased for increase in area values from .4 to 1.0. On an average, a 40% increase in the detection rates was observed for the three statistics when the area value increased from .4 to 1.0. In practical settings, if the size of DIF in an item is very small, it is likely to go undetected especially when the sample sizes are small. However, in such cases, it can be argued that the size of DIF is not substantial enough to be of any practical significance.

The most significant result in the study concerned the impact of the type of item on the DIF detection rates. Detection rates were highest for high discrimination items followed by moderate and low discriminating items. Detection rates were lowest for high difficulty items followed by items of moderate difficulty and low difficulty. Highly difficult items are likely to be answered correctly by only a limited number of examinees at the extreme end of the ability continuum and therefore function differently for a relatively small number of examinees. Because very difficult items are not very common in standardized tests, they finding may not be a matter of great concern in practice.

The proportion of items containing DIF appears to have minimal effect on the DIF detection rates (at least for the percentages

simulated in the study and when two-stage procedures are used). This may be due to the two-stage procedure adopted in computing the SIB and MH statistics. Items identified as DIF in the first computations were removed when forming the score groups for computing the DIF statistics for the second time. Because the proportion of items showing DIF did not affect the DIF detection rates, it is recommended that the two-stage procedure should be implemented in practice when computing the DIF statistics.

In conclusion, the results of this study show that in general, both the MH and the SIB procedures are viable methods for detecting uniform DIF in test items although both have their advantages as well as their limitations.

Table 4.1

Item Parameters Used to Generate Items with DIF for the Distribution and the Power Studies

Item No.	Type of Item		DIF Effect Size	Ref. b1	Foc. b2	Ref. a1	Foc. a2
Distribution Study							
1.	Low b	Low a		-1.50	-1.50	0.50	0.50
2.	Low b	Medium a		-1.50	-1.50	1.00	1.00
3.	Low b	High a		-1.50	-1.50	1.50	1.50
4.	Medium b	Low a		0.00	0.00	0.50	0.50
5.	Medium b	Medium a		0.00	0.00	1.00	1.00
6.	Medium b	High a		0.00	0.00	1.50	1.50
7.	High b	Low a		1.50	1.50	0.50	0.50
8.	High b	Medium a		1.50	1.50	1.00	1.00
9.	High b	High a		1.50	1.50	1.50	1.50
DIF Study							
1.	Low b	Medium a	.4	-1.80	-1.28	1.00	1.00
2.			.6	-1.92	-1.14	1.00	1.00
3.			.8	-2.04	-1.01	1.00	1.00
4.			1.0	-2.16	-0.88	1.00	1.00
5.	Low b	High a	.4	-1.80	-1.28	1.50	1.50
6.			.6	-1.92	-1.14	1.50	1.50
7.			.8	-2.04	-1.01	1.50	1.50
8.			1.0	-2.16	-0.88	1.50	1.50
9.	Medium b	Low a	.4	-0.26	0.26	0.50	0.50
10			.6	-0.39	0.39	0.50	0.50
11.			.8	-0.51	0.51	0.50	0.50
12.			1.0	-0.64	0.64	0.50	0.50
13.	Medium b	High a	.4	-0.26	0.26	1.50	1.50
14.			.6	-0.39	0.39	1.50	1.50
15.			.8	-0.51	0.51	1.50	1.50
16.			1.0	-0.64	0.64	1.50	1.50
17.	High b	Low a	.4	1.28	1.80	0.50	0.50
18.			.6	1.14	1.92	0.50	0.50
19.			.8	1.01	2.04	0.50	0.50
20.			1.0	0.88	2.16	0.50	0.50
21.	High b	Medium a	.4	1.28	1.80	1.00	1.00
22.			.6	1.14	1.92	1.00	1.00
23.			.8	1.01	1.24	1.00	1.00
24.			1.0	0.88	2.16	1.00	1.00

Table 4.2

Kolmogorov-Smirnov and Wilks-Shapiro Test Results for Testing the Distributional Assumptions of the Simultaneous Item Bias Test Statistic

Sample Size		Item Parameter							
Ref.	Foc.	b	a	Mean	S.D.	K-S	p	W-S	p
300	100	-1.50	0.50	-0.02	1.02	.015	.98	.989	.88
		-1.50	1.00	-0.08	1.07	.036	.16	.986	.37
		-1.50	1.50	-0.11	1.09	.044	.04*	.981	.00*
		0.00	0.50	0.00	0.97	.018	.90	.986	.39
		0.00	1.00	-0.03	1.03	.030	.32	.987	.55
		0.00	1.50	-0.04	0.99	.032	.27	.989	.80
		1.50	0.50	0.02	1.04	.021	.75	.988	.78
		1.50	1.00	0.06	1.05	.047	.03*	.988	.72
		1.50	1.50	0.07	1.07	.035	.17	.990	.96
300	200	-1.50	0.50	-0.02	0.98	.027	.46	.988	.84
		-1.50	1.00	-0.05	0.96	.039	.10	.986	.44
		-1.50	1.50	-0.03	0.97	.024	.60	.985	.31
		0.00	0.50	-0.03	1.03	.021	.79	.989	.81
		0.00	1.00	0.02	0.99	.027	.46	.990	.93
		0.00	1.50	0.02	1.00	.024	.61	.988	.76
		1.50	0.50	-0.03	1.01	.052	.00*	.985	.28
		1.50	1.00	-0.03	0.97	.032	.26	.986	.44
		1.50	1.50	-0.07	1.02	.027	.48*	.984	.09
300	300	-1.50	0.50	0.02	0.96	.036	.15	.987	.67
		-1.50	1.00	0.00	0.99	.022	.69	.989	.84
		-1.50	1.50	-0.01	0.99	.018	.91	.990	.92
		0.00	0.00	0.00	1.05	.027	.45	.990	.94
		0.00	1.00	0.00	1.06	.033	.22	.986	.33
		0.00	1.50	-0.02	1.00	.022	.73	.988	.71
		1.50	0.50	0.01	1.01	.030	.31	.985	.16
		1.50	1.00	0.01	0.99	.021	.77	.989	.87
		1.50	1.50	-0.06	0.99	.037	.14	.985	.27

Continued, next page.

Table 4.2--continued:

Sample Size		Item Parameter							
Ref.	Foc.	b	a	Mean	S.D.	K-S	p	W-S	p
500	100	-1.50	0.50	-0.08	1.07	.041	.06	.985	.30
		-1.50	1.00	-0.09	1.07	.040	.09	.987	.53
		-1.50	1.50	-0.20	1.16	.060	.00	.984	.08
		0.00	0.50	-0.04	1.00	.037	.14	.985	.23
		0.00	1.00	0.01	1.04	.022	.68	.984	.12
		0.00	1.50	0.03	1.04	.029	.35	.988	.69
		1.50	0.50	0.04	1.06	.040	.10	.987	.55
		1.50	1.00	0.02	0.98	.019	.87	.988	.67
		1.50	1.50	0.12	1.05	.044	.04*	.980	.00*
500	200	-1.50	0.50	-0.06	1.00	.033	.22	.987	.64
		-1.50	1.00	0.00	1.00	.024	.59	.988	.79
		-1.50	1.50	-0.08	1.05	.029	.38	.987	.50
		0.00	0.50	0.00	1.04	.022	.70	.985	.29
		0.00	1.00	-0.04	1.00	.032	.26	.988	.68
		0.00	1.50	-0.04	1.02	.031	.28	.989	.85
		1.50	0.50	-0.01	0.99	.020	.84	.989	.84
		1.50	1.00	0.06	0.98	.032	.26	.986	.38
		1.50	1.50	0.00	1.01	.015	.96	.986	.39
500	300	-1.50	0.50	-0.06	0.98	.040	.08	.984	.09
		-1.50	1.00	-0.03	1.01	.024	.64	.986	.38
		-1.50	1.50	-0.02	1.00	.023	.64	.985	.27
		0.00	0.50	0.03	1.03	.019	.85	.984	.09
		0.00	1.00	-0.03	1.01	.045	.03*	.991	.98
		0.00	1.50	0.05	1.02	.038	.10	.987	.57
		1.50	0.50	-0.03	0.99	.023	.69	.985	.28
		1.50	1.00	0.03	0.99	.027	.44	.987	.49
		1.50	1.50	0.03	1.03	.030	.32	.989	.55

Continued, next page.

Table 4.2--continued:

Sample Size		Item Parameter							
Ref.	Foc.	b	a	Mean	S.D.	K-S	p	W-S	p
1000	100	-1.50	0.50	0.05	1.04	.030	.21	.989	.70
		-1.50	1.00	0.06	1.02	.037	.12	.985	.26
		-1.50	1.50	-0.25	1.21	.077	.00*	.939	.00*
		0.00	0.50	-0.06	1.03	.057	.00*	.989	.90
		0.00	1.00	-0.05	1.04	.035	.19	.987	.53
		0.00	1.50	-0.01	1.03	.019	.87	.990	.93
		1.50	0.50	-0.13	1.08	.050	.01*	.981	.00*
		1.50	1.00	-0.16	1.16	.058	.00*	.981	.01*
		1.50	1.50	0.11	1.09	.056	.00*	.984	.20
1000	200	-1.50	0.50	-0.01	0.95	.019	.20	.989	.84
		-1.50	1.00	-0.06	1.05	.038	.11	.987	.50
		-1.50	1.50	-0.11	1.05	.042	.05	.982	.02*
		0.00	0.50	-0.01	1.01	.018	.90	.987	.50
		0.00	1.00	0.04	1.02	.041	.07	.986	.37
		0.00	1.50	0.00	1.03	.023	.66	.986	.38
		1.50	0.50	0.04	1.01	.021	.79	.987	.59
		1.50	1.00	0.04	1.01	.039	.06	.984	.08
		1.50	1.50	0.05	0.98	.034	.20	.983	.04*
1000	300	-1.50	0.50	-0.01	1.00	.016	.95	.988	.66
		-1.50	1.00	0.00	0.98	.022	.72	.985	.34
		-1.50	1.50	0.11	1.00	.050	.01	.985	.21
		0.00	0.50	0.03	1.05	.036	.15	.987	.55
		0.00	1.00	-0.01	1.03	.023	.68	.986	.37
		0.00	1.50	-0.02	1.00	.020	.82	.990	.92
		1.50	0.50	0.00	0.96	.028	.42	.986	.36
		1.50	1.00	0.01	1.02	.015	.99	.986	.37
		1.50	1.50	0.03	1.01	.024	.61	.981	.41

Critical value for K-S test statistic at .05 level (1000 replications) = 0.043
Critical value for W-S test statistic at .05 level of significance = 0.983

Table 4.3

Kolmogorov-Smirnov Test Results for Testing the Distributional Assumptions of the Mantel-Haenszel(1) and Mantel-Haenszel(2) Statistics

Sample Size		Item Parameter		Mantel-Haenszel(1)				Mantel-Haenszel(2)			
Ref.	Foc.	b	a	Mean	S.D.	K-S	p	Mean	S.D.	K-S	p
300	100	-1.50	0.50	0.74	1.18	.143	.00*	0.97	1.35	.056	.00*
		-1.50	1.00	0.75	1.25	.160	.00*	1.05	1.49	.058	.00*
		-1.50	1.50	0.67	1.08	.169	.00*	1.03	1.36	.065	.00*
		0.00	0.50	0.74	1.09	.122	.00*	0.93	1.23	.055	.00*
		0.00	1.00	0.79	1.23	.139	.00*	0.99	1.39	.058	.00*
		0.00	1.50	0.74	1.15	.143	.00*	0.95	1.31	.044	.04*
		1.50	0.50	0.86	1.35	.131	.00*	1.05	1.50	.067	.00*
		1.50	1.00	0.81	1.24	.095	.00*	1.01	1.39	.059	.00*
		1.50	1.50	0.81	1.24	.122	.00*	1.02	1.41	.052	.00*
300	200	-1.50	0.50	0.83	1.32	.123	.00*	1.01	1.47	.065	.00*
		-1.50	1.00	0.70	1.08	.148	.00*	0.93	1.26	.056	.00*
		-1.50	1.50	0.67	1.07	.161	.00*	0.94	1.29	.046	.03*
		0.00	0.50	0.87	1.38	.103	.00*	1.02	1.49	.048	.02*
		0.00	1.00	0.80	1.19	.115	.00*	0.96	1.31	.060	.00*
		0.00	1.50	0.80	1.20	.104	.00*	0.97	1.33	.061	.00*
		1.50	0.50	0.83	1.24	.148	.00*	0.98	1.36	.076	.00*
		1.50	1.00	0.81	1.21	.117	.00*	0.96	1.33	.053	.00*
		1.50	1.50	0.81	1.23	.129	.00*	0.97	1.36	.060	.00*
300	300	-1.50	0.50	0.75	1.16	.130	.00*	0.90	1.28	.054	.00*
		-1.50	1.00	0.75	1.22	.148	.00*	0.95	1.38	.048	.02*
		-1.50	1.50	0.73	1.23	.156	.00*	0.97	1.42	.066	.00*
		0.00	0.50	0.97	1.54	.107	.00*	1.11	1.65	.045	.03*
		0.00	1.00	0.95	1.34	.101	.00*	1.11	1.46	.054	.00*
		0.00	1.50	0.84	1.27	.110	.00*	0.99	1.38	.044	.04*
		1.50	0.50	0.89	1.34	.104	.00*	1.03	1.44	.054	.00*
		1.50	1.00	0.80	1.31	.145	.00*	0.93	1.42	.065	.00*
		1.50	1.50	0.85	1.29	.095	.00*	1.00	1.40	.052	.00*

Continued, next page.

Table 4.3--continued:

Sample Size		Item Parameter		Mantel-Haenszel(1)				Mantel-Haenszel(2)			
Ref.	Foc.	b	a	Mean	S.D.	K-S	p	Mean	S.D.	K-S	p
500	100	-1.50	0.50	0.80	1.28	.126	.00*	1.01	1.45	.044	.04*
		-1.50	1.00	0.83	1.20	.163	.00*	1.01	1.43	.059	.00*
		-1.50	1.50	0.67	1.06	.179	.00*	0.99	1.33	.053	.00*
		0.00	0.50	0.78	1.19	.130	.00*	0.95	1.33	.064	.00*
		0.00	1.00	0.85	1.30	.105	.00*	1.05	1.45	.059	.00*
		0.00	1.50	0.82	1.26	.104	.00*	1.03	1.42	.048	.02*
		1.50	0.50	0.86	1.26	.090	.00*	1.04	1.39	.044	.04*
		1.50	1.00	0.74	1.15	.153	.00*	0.92	1.29	.064	.00*
		1.50	1.50	0.81	1.21	.104	.00*	1.01	1.36	.069	.00*
500	200	-1.50	0.50	0.80	1.27	.113	.00*	0.96	1.40	.056	.00*
		-1.50	1.00	0.82	1.33	.126	.00*	1.04	1.50	.050	.01*
		-1.50	1.50	0.76	1.23	.134	.00*	1.02	1.43	.051	.00*
		0.00	0.50	0.90	1.41	.111	.00*	1.04	1.52	.052	.00*
		0.00	1.00	0.86	1.36	.110	.00*	1.00	1.47	.063	.00*
		0.00	1.50	0.85	1.32	.113	.00*	1.00	1.44	.067	.00*
		1.50	0.50	0.87	1.32	.110	.00*	1.00	1.42	.045	.03*
		1.50	1.00	0.80	1.20	.110	.00*	0.95	1.30	.052	.00*
		1.50	1.50	0.81	1.23	.113	.00*	0.96	1.34	.056	.00*
500	300	-1.50	0.50	0.78	1.16	.117	.00*	0.92	1.27	.060	.00*
		-1.50	1.00	0.78	1.16	.106	.00*	0.97	1.30	.052	.00*
		-1.50	1.50	0.77	1.24	.148	.00*	0.99	1.41	.053	.00*
		0.00	0.50	0.93	1.39	.098	.00*	1.05	1.49	.054	.00*
		0.00	1.00	0.88	1.41	.102	.00*	1.01	1.50	.053	.00*
		0.00	1.50	0.89	1.35	.102	.00*	1.03	1.45	.051	.00*
		1.50	0.50	0.86	1.29	.113	.00*	0.98	1.38	.056	.00*
		1.50	1.00	0.84	1.29	.124	.00*	0.96	1.39	.056	.00*
		1.50	1.50	0.91	1.37	.097	.00*	1.05	1.47	.052	.00*

Continued, next page.

Table 4.3--continued:

Sample Size		Item Parameter		Mantel-Haenszel(1)				Mantel-Haenszel(2)			
Ref.	Foc.	b	a	Mean	S.D.	K-S	p	Mean	S.D.	K-S	p
1000	100	-1.50	0.50	0.82	1.27	.118	.00*	1.17	1.98	.062	.00*
		-1.50	1.00	0.77	1.26	.144	.00*	1.34	2.55	.059	.00*
		-1.50	1.50	0.66	1.07	.168	.00*	1.61	3.23	.063	.00*
		0.00	0.50	0.85	1.48	.120	.00*	1.07	1.74	.062	.00*
		0.00	1.00	0.80	1.28	.146	.00*	1.09	1.63	.068	.00*
		0.00	1.50	0.79	1.24	.114	.00*	1.06	1.60	.065	.00*
		1.50	0.50	0.84	1.42	.126	.00*	1.08	1.74	.061	.00*
		1.50	1.00	0.78	1.22	.113	.00*	1.04	1.61	.069	.00*
		1.50	1.50	0.91	1.42	.111	.00*	1.12	1.77	.067	.00*
1000	200	-1.50	0.50	0.76	1.10	.120	.00*	0.91	1.21	.056	.00*
		-1.50	1.00	0.84	1.26	.129	.00*	1.04	1.41	.057	.00*
		-1.50	1.50	0.76	1.19	.136	.00*	0.99	1.38	.065	.00*
		0.00	0.50	0.90	1.32	.128	.00*	1.03	1.42	.063	.00*
		0.00	1.00	0.84	1.31	.133	.00*	0.98	1.42	.058	.00*
		0.00	1.50	0.84	1.26	.103	.00*	0.99	1.37	.055	.00*
		1.50	0.50	0.85	1.32	.132	.00*	0.97	1.41	.053	.00*
		1.50	1.00	0.88	1.28	.111	.00*	1.02	1.38	.050	.01*
		1.50	1.50	0.82	1.27	.102	.00*	0.96	1.37	.061	.00*
1000	300	-1.50	0.50	0.87	1.28	.097	.00*	1.01	1.38	.063	.00*
		-1.50	1.00	0.79	1.22	.124	.00*	0.95	1.35	.058	.00*
		-1.50	1.50	0.73	1.15	.147	.00*	0.92	1.30	.065	.00*
		0.00	0.50	0.95	1.45	.116	.00*	1.07	1.54	.055	.00*
		0.00	1.00	0.91	1.35	.102	.00*	1.04	1.44	.058	.00*
		0.00	1.50	0.89	1.38	.093	.00*	1.01	1.47	.044	.04*
		1.50	0.50	0.80	1.18	.123	.00*	0.90	1.26	.067	.00*
		1.50	1.00	0.89	1.37	.117	.00*	1.00	1.46	.059	.00*
		1.50	1.50	0.89	1.29	.102	.00*	1.01	1.37	.052	.00*

Critical value for K-S test statistic at .05 level (1000 replications) = 0.043

Table 4.4

Mean Percent Type I Error Rates of the Simultaneous Item Bias, Mantel-Haenszel(1) and Mantel-Haenszel(2) Statistics (Distribution Study)

Sample Size		Item Parameter		SIB Statistic		MH1 Statistic		MH2 Statistic	
Ref.	Foc.	b	a	α=.05	α=.01	α=.05	α=.01	α=.05	α=.01
				%	%	%	%	%	%
300	100	-1.50	0.50	6.0	1.2	3.5	0.5	4.4	1.3
		-1.50	1.00	7.3	2.5	3.4	0.6	5.1	1.1
		-1.50	1.50	7.0	2.2	2.3	0.3	4.8	0.9
		0.00	0.50	4.4	0.4	2.6	0.3	3.5	0.5
		0.00	1.00	5.7	1.2	4.3	0.5	5.3	1.1
		0.00	1.50	4.5	0.7	3.0	0.5	4.6	0.9
		1.50	0.50	6.1	1.0	4.2	0.8	5.6	1.3
		1.50	1.00	5.3	1.5	3.5	0.6	4.7	1.0
		1.50	1.50	7.1	1.6	3.6	0.5	4.8	0.8
300	200	-1.50	0.50	5.1	1.5	3.5	0.8	4.4	0.8
		-1.50	1.00	5.0	0.6	3.3	0.1	4.1	0.6
		-1.50	1.50	4.6	0.7	2.6	0.3	4.1	0.8
		0.00	0.50	5.7	1.3	4.1	0.6	5.4	1.0
		0.00	1.00	4.4	0.9	3.1	0.5	5.0	0.6
		0.00	1.50	5.4	1.0	3.4	0.5	4.2	0.7
		1.50	0.50	6.1	1.0	3.8	0.4	5.0	0.9
		1.50	1.00	4.1	1.1	2.5	0.9	3.2	1.1
		1.50	1.50	5.2	0.6	3.8	0.7	4.8	0.8
300	300	-1.50	0.50	4.8	0.6	3.4	0.2	4.6	0.7
		-1.50	1.00	4.5	1.0	3.5	0.4	4.6	1.1
		-1.50	1.50	5.0	0.9	2.2	1.1	3.8	1.2
		0.00	0.50	5.6	1.2	4.7	1.1	6.3	1.5
		0.00	1.00	5.7	1.2	5.5	0.6	6.6	1.0
		0.00	0.50	4.7	1.0	3.5	0.4	4.9	0.6
		1.50	1.00	5.3	1.6	4.1	0.8	5.1	1.1
		1.50	1.50	5.2	1.6	4.2	0.9	4.8	1.1
		1.50	1.50	4.4	0.4	5.6	1.0	5.1	0.8

Continued, next page.

Table 4.4--continued:

Sample Size		Item Parameter		SIB Statistic		MH1 Statistic		MH2 Statistic	
Ref.	Foc.	b	a	α=.05	α=.01	α=.05	α=.01	α=.05	α=.01
				%	%	%	%	%	%
500	100	-1.50	0.50	6.5	2.4	4.4	1.1	5.5	1.4
		-1.50	1.00	6.5	2.6	2.9	0.9	5.2	1.1
		-1.50	1.50	7.8	2.5	2.4	0.2	5.0	0.7
		0.00	0.50	5.6	1.8	3.4	0.4	4.4	0.9
		0.00	1.00	6.0	1.4	4.6	0.5	6.2	0.7
		0.00	1.50	6.2	1.1	3.6	0.7	5.2	1.3
		1.50	0.50	5.5	1.4	4.3	0.5	5.4	0.8
		1.50	1.00	4.2	0.6	2.7	0.6	4.3	0.8
		1.50	1.50	5.6	1.6	3.2	0.7	4.4	0.9
500	200	-1.50	0.50	5.5	1.0	3.5	0.8	3.9	1.0
		-1.50	1.00	4.8	1.4	4.0	0.9	5.6	1.5
		-1.50	1.50	6.1	1.5	3.5	0.7	5.7	1.0
		0.00	0.50	5.3	1.6	4.0	1.0	5.8	1.3
		0.00	1.00	4.9	1.3	4.2	1.2	5.4	1.0
		0.00	0.50	5.4	0.9	4.2	0.7	5.4	1.0
		1.50	1.00	5.0	1.3	3.8	0.8	4.4	1.8
		1.50	1.50	4.8	0.8	3.2	0.6	4.0	0.6
		1.50	1.50	5.1	0.8	3.7	0.8	4.8	1.0
500	300	-1.50	0.50	4.8	0.8	3.4	0.3	4.4	0.4
		-1.50	1.00	4.9	0.7	2.8	0.4	3.9	0.7
		-1.50	1.50	4.9	1.0	3.9	0.7	5.0	1.0
		0.00	0.50	5.7	1.3	4.9	1.0	6.3	1.0
		0.00	1.00	5.9	1.0	5.1	0.7	5.7	0.8
		0.00	1.50	5.2	0.9	4.6	0.8	5.8	1.1
		1.50	0.50	4.8	1.0	4.2	0.7	5.3	0.8
		1.50	1.00	4.7	0.8	4.1	0.7	5.2	0.8
		1.50	1.50	6.2	1.3	4.9	0.7	6.0	0.9

Continued, next page.

Table 4.4--continued:

Sample Size		Item Parameter		SIB Statistic		MH1 Statistic		MH2 Statistic	
Ref.	Foc.	b	a	α=.05	α=.01	α=.05	α=.01	α=.05	α=.01
				%	%	%	%	%	%
1000	100	-1.50	0.50	7.2	1.5	3.4	0.9	5.5	1.1
		-1.50	1.00	7.8	2.7	3.4	0.7	5.7	1.1
		-1.50	1.50	8.2	2.9	2.2	0.3	4.3	0.7
		0.00	0.50	5.5	1.9	4.3	1.1	5.1	1.5
		0.00	1.00	7.1	1.6	4.1	0.9	5.5	1.0
		0.00	1.50	6.0	1.3	3.3	0.6	4.5	0.7
		1.50	0.50	6.4	1.6	3.7	1.2	4.9	1.4
		1.50	1.00	6.1	1.9	4.0	0.6	4.9	1.0
		1.50	1.50	7.6	2.3	4.4	0.9	5.6	1.3
1000	200	-1.50	0.50	3.4	0.5	2.4	0.3	3.3	0.4
		-1.50	1.00	6.3	1.5	4.0	0.9	5.1	1.1
		-1.50	1.50	6.7	2.0	3.5	0.3	5.2	0.9
		0.00	0.50	5.3	1.0	4.3	0.7	5.3	0.7
		0.00	1.00	5.2	1.4	3.5	0.9	5.0	1.5
		0.00	1.50	5.5	0.9	3.2	0.9	4.1	1.2
		1.50	0.50	5.4	1.1	4.3	0.9	4.6	1.0
		1.50	1.00	5.1	1.0	4.1	0.6	4.5	1.1
		1.50	1.50	5.1	1.2	3.8	0.8	4.7	1.2
1000	300	-1.50	0.50	5.4	0.9	4.4	0.6	5.5	0.6
		-1.50	1.00	5.5	1.1	3.7	0.2	5.1	0.6
		-1.50	1.50	5.8	1.1	3.1	0.4	4.4	0.5
		0.00	0.50	6.4	1.1	4.9	1.0	5.6	1.4
		0.00	1.00	5.4	1.2	4.4	1.2	4.9	1.2
		0.00	1.50	5.8	1.0	4.1	0.9	5.0	1.1
		1.50	0.50	4.3	0.7	3.3	0.5	4.3	0.5
		1.50	1.00	5.5	1.5	3.7	1.0	5.0	1.2
		1.50	1.50	5.4	0.7	3.7	0.7	4.4	0.8

Table 4.5

Mean Percent Detection Rates of the Simultaneous Item Bias, Mantel-Haenszel(1) and Mantel-Haenszel(2) Statistics Under all Conditions (Power Study)

Factor		SIB Statistic α=.05 (%)	SIB Statistic α=.01 (%)	MH(1) Statistic α=.05 (%)	MH(1) Statistic α=.01 (%)	MH(2) Statistic α=.05 (%)	MH(2) Statistic α=.01 (%)
Sample Size							
Ref.	Foc.						
300	100	64	49	60	46	63	51
300	200	77	65	72	61	75	63
300	300	83	74	78	68	80	71
500	100	63	49	60	47	63	48
500	200	79	67	73	62	77	65
500	300	84	75	79	69		
1000	100	63	48	60	48	63	48
1000	200	77	66	72	61	76	64
1000	300	83	77	73	68	81	70
Test Length							
40		76	65	70	59	73	61
60		74	61	69	58	73	61
Ability Distribution							
Equal		77	63	76	63	77	63
Unequal		74	63	64	54	69	57
Proportion of DIF							
10%		75	63	71	60	73	62
20%		74	62	68	58	70	60
Type of Item							
Low b	Medium a	88	79	89	81	91	82
Low b	High a	92	85	93	88	94	89
Med b	Low a	72	57	67	51	70	54
Med b	High a	94	87	93	86	95	87
High b	Low a	50	33	39	23	44	27
High b	Medium a	53	37	41	26	46	30
DIF Effect Size							
.4		52	36	47	32	50	34
.6		73	59	67	55	71	57
.8		84	74	79	69	82	73
1.0		91	83	86	79	89	92

Table 4.6

Mean Percent Type I Error Rates of the Simultaneous Item Bias, Mantel-Haenszel(1) and Mantel-Haenszel(2) Statistics for the Non-DIF Test Items (Power Study)

	SIB Statistic		MH(1) Statistic		MH(2) Statistic	
	α=.05	α=.01	α=.05	α=.01	α=.05	α=.01
Factor	(%)	(%)	(%)	(%)	(%)	(%)
Sample Size						
Ref. Foc.						
300 100	6.6	1.7	3.9	0.7	5.2	1.0
300 200	6.5	1.6	4.2	0.8	5.7	1.3
300 300	6.9	1.8	4.5	1.0	5.7	1.2
500 100	6.8	2.0	4.0	0.8	5.2	1.0
500 200	6.6	1.6	4.5	1.0	5.7	1.3
500 300	7.1	1.9	4.9	0.8	5.9	1.1
1000 100	7.0	2.1	4.0	1.0	5.2	1.3
1000 200	6.9	1.9	4.8	1.1	5.9	1.5
1000 300	7.1	1.8	5.0	1.1	6.1	1.6
Test Length						
40	7.0	1.8	4.5	0.9	5.6	1.2
60	6.6	1.7	4.4	0.9	5.5	1.2
Ability Distribution						
Equal	6.0	1.5	4.1	0.8	5.1	1.1
Unequal	7.8	2.1	4.8	1.0	5.6	1.2

Table 4.7

Mean Percent Detection Rates of Mantel-Haenszel(1) and Mantel-Haenszel(2) Statistics for Equal Ability Distribution for Different Sample Sizes and Types of Item

		High b, Low a				High b, Medium a			
		MH(1)		MH(2)		MH(1)		MH(2)	
		α=.05	α=.01	α=.05	α=.01	α=.05	α=.01	α=.05	α=.01
Factor		(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
Sample Size									
Ref.	Foc.								
300	100	37	19	40	23	42	23	48	25
300	200	51	33	55	36	61	43	64	45
300	300	63	43	65	46	72	58	73	60
500	100	38	20	43	24	48	27	55	30
500	200	57	37	60	41	72	55	74	58
500	300	70	49	71	52	75	64	77	65
1000	100	46	25	50	28	49	33	53	36
1000	200	63	44	66	47	72	58	73	60
1000	300	75	59	77	61	84	70	85	72

Table 4.8

Mean Percent Detection Rates of Mantel-Haenszel(1) and Mantel-Haenszel(2) Statistics for Unequal Ability Distribution for Different Sample Sizes and Types of Item

		High b, Low a				High b, Medium a			
		MH(1)		MH(2)		MH(1)		MH(2)	
		α=.05	α=.01	α=.05	α=.01	α=.05	α=.01	α=.05	α=.01
Factor		(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
Sample Size									
Ref.	Foc.								
300	100	15	6	19	7	12	4	15	6
300	200	20	9	22	10	19	7	21	9
300	300	33	19	37	22	72	58	73	60
500	100	16	4	19	5	48	27	55	30
500	200	27	14	30	14	72	55	74	58
500	300	41	19	44	21	75	64	77	65
1000	100	16	7	18	7	49	33	53	36
1000	200	28	13	31	14	72	58	73	60
1000	300	29	14	31	16	84	70	85	72

Table 4.9

Mean Percent Type I Error Rates of Mantel-Haenszel(1) and Mantel-Haenszel(2) Statistics for Unequal Ability Distribution for Different Sample Sizes and Types of Item

		High b, Low a				High b, Medium a			
		MH(1)		MH(2)		MH(1)		MH(2)	
		α=.05	α=.01	α=.05	α=.01	α=.05	α=.01	α=.05	α=.01
Factor		(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
Sample Size									
Ref.	Foc.								
300	100	3.8	0.5	5.1	0.9	3.7	0.6	5.4	0.9
300	200	4.1	0.9	5.4	1.3	4.4	0.6	5.7	0.9
300	300	4.1	0.6	5.1	0.9	4.0	0.8	5.2	1.1
500	100	4.3	0.8	5.5	1.0	3.7	0.7	4.7	1.1
500	200	4.0	0.7	5.2	1.1	4.0	0.6	5.0	0.9
500	300	3.8	0.7	4.9	0.9	4.0	0.8	5.0	1.1
1000	100	4.1	1.0	5.6	1.3	3.4	0.8	4.8	1.1
1000	200	3.5	0.7	4.9	1.0	3.6	0.8	4.6	1.0
1000	300	4.2	0.8	4.8	1.0	5.0	0.8	5.6	0.9

Table 4.10

Mean Percent Type I Error Rates of Mantel-Haenszel(1) and Mantel-Haenszel(2) Statistics for Unequal Ability Distribution for Different Sample Sizes and Types of Item

		High b, Low a				High b, Medium a			
		MH(1)		MH(2)		MH(1)		MH(2)	
		α=.05	α=.01	α=.05	α=.01	α=.05	α=.01	α=.05	α=.01
Factor		(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
Sample Size									
Ref.	Foc.								
300	100	3.9	0.9	5.6	1.2	4.1	0.9	6.1	1.3
300	200	3.9	0.7	5.4	1.0	4.8	1.2	6.6	1.6
300	300	4.9	1.0	6.2	1.4	5.1	1.1	6.0	1.5
500	100	4.2	0.8	5.9	1.3	4.6	0.9	6.4	1.4
500	200	4.7	0.8	6.3	1.2	5.3	1.4	6.4	1.6
500	300	5.1	1.4	6.3	1.8	5.2	1.4	6.4	1.7
1000	100	4.0	1.2	5.4	1.5	4.2	0.9	5.6	1.3
1000	200	5.7	1.5	6.6	1.8	5.4	1.4	6.8	1.7
1000	300	6.1	1.6	7.0	1.9	6.6	1.8	7.7	2.1

C H A P T E R V

IDENTIFICATION OF ITEMS THAT SHOW NON-UNIFORM DIF

Introduction

In recent years, there has been a great deal of concern over the issue of differential item functioning (DIF) in educational data. DIF is said to exist if examinees having the same underlying ability have different probabilities of getting an item correct regardless of group membership. From an item response theory (IRT) perspective, an item shows DIF if the item characteristic curves (ICCs) evaluated across two different subgroups are not identical.

According to Mellenbergh (1982), two types of DIF can occur in educational dichotomous data. Uniform DIF is said to occur when there is no interaction between the ability level and group membership. Non-uniform DIF is said to occur when there is interaction between the ability level and group membership. In terms of IRT, uniform and non-uniform DIF are represented by parallel and non-parallel ICCs respectively. In general, although uniform DIF is seen more often than non-uniform DIF in standardized tests, identification of non-uniformly functioning items in real data have been reported in previous research (Mellenbergh, 1982; Hambleton & Rogers 1989; Linn et al. 1981).

This study will investigate the detection of non-uniform DIF using three popular statistical DIF detection procedures: the Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988), the Simultaneous Item Bias (SIBTEST, henceforth referred to as SIB) procedure (Shealy & Stout, 1993) and the Logistic Regression (LR) procedure (Swaminathan &

Rogers, 1990). Both the MH and the SIB procedures are non-parametric, computationally simple, easy to implement in practice and provide statistics that have associated tests of significance.

Swaminathan and Rogers (1990) have shown that the LR procedure despite its parametric nature, can be easily implemented in practice. A major advantage of the LR procedure is that it is a model-based procedure with the ability variable treated as continuous. It also allows for testing the hypothesis of no interaction between the ability variable and the group variable. In fact, the MH procedure can be conceptualized as being based on the LR model where the ability variable is treated as discrete and no interaction between the ability variable and group membership is permitted. The LR procedure would therefore be expected to improve on the MH procedure for detecting non-uniform DIF.

Previous research studies have shown that the MH, SIB and the LR procedures are equally effective in the identification of uniform DIF (Swaminathan & Rogers, 1990; Rogers & Swaminathan, 1993; Ackerman 1992; Roussos & Stout, 1993; Narayanan & Swaminathan, 1993). Rogers (1989) and Rogers and Swaminathan (1993) using simulated data showed that the MH procedure is not capable of identifying disordinal non-uniform DIF where the ICCs for the two groups cross near the middle of the examinee score distribution (i.e., items of average difficulty). Such items can be adequately identified with the logistic regression procedure that includes a term for interaction between group membership and ability. The major advantage of the LR procedure is that it can be expanded to condition on more than one ability variable.

Research Objectives

Recently, a modification of the SIB procedure, known as CROSSING-SIBTEST (CRO-SIB), has been developed by Li and Stout (1993). It is designed to detect non-uniform DIF and has the potential for conditioning on more than one ability variable. Because of its newness, the CRO-SIB procedure has not been extensively studied. Given the possibility that it could be superior to the MH and LR procedures in some situations, a detailed investigation of the three procedures is timely.

The main purposes for conducting this study are: (1) to investigate and compare the power and Type I error rates of the MH, SIB, and LR procedures, and (2) to determine the conditions under which each procedure is optimal for detecting non-uniform DIF.

Method

Description of the Power Study

This research study was conducted on simulated data sets. A number of factors affecting the DIF detection rates have been identified in previous research (Rogers, 1989; Rogers & Swaminathan, 1993; Mazor et al., 1992) In this study, five such factors were manipulated: sample size, proportion of items containing DIF, ability distribution differences, DIF effect size and type of item. The two reference group sample sizes (500, 1000) were crossed with the two focal group sample sizes (200, 500) to produce four conditions related to sample size. The study was confined to a single test length of 40 items which is approximately the average length of standardized achievement subtests.

The impact of the ability distribution differences between the reference and the focal groups was investigated by examining two different conditions. In the first condition, the ability distributions for the two groups were set to be equal with mean 0.0 and standard deviation equal to one. In the second condition, the mean was set to 0.0 and -1.0 for the reference and the focal groups respectively, again with both standard deviations set equal to one. Distributions that differ by one standard deviation was chosen to simulate the cases sometimes found in DIF studies (for example, Hambleton & Rogers, 1989). Since the percent of DIF items can contaminate the conditioning variable, the proportion of items containing DIF was set at three levels: 0%, 10% and 20%.

The size of DIF was manipulated using the area between the ICCs for the two groups as the measure of DIF effect size. Four levels of DIF effect size corresponding to area values of .4, .6, .8 and 1.0 were chosen to reflect DIF effect size values ranging from a small amount of DIF to a fairly large amount of DIF. Non-uniform DIF was simulated by keeping the b-parameters for the two groups the same, while varying the a-parameters for the two groups. 16 items showing non-uniform DIF were simulated by varying the level of the common difficulty level (low, medium, high), the level of the discrimination parameter for the two groups (low, medium, high) and DIF effect size (area values of .4, .6, .8 and 1.0). In all, four types of item were studied: (1) low b, high a; (2) medium b, low a; (3) medium b, high a; and (4) high b, low a.

To simulate a 40-item test with 10% of the items showing DIF (i.e., four items) and to accommodate the characteristics of items

that may affect DIF detection, it was necessary to distribute the 16 items into four 40-item tests. Similarly, to simulate 20% of the items showing DIF (i.e., eight items), the 16 DIF items were distributed into two 40-item tests. Item parameter values for the non-DIF items were kept the same in all the 40-item tests. They were randomly chosen from published item parameter values from an administration of the Graduate Management Admission Test (Kingston, Leary & Wightman, 1988). The c-parameters for all the items were set equal to .20.

Data for the study were simulated according to the three-parameter model using the program DATAGEN (Hambleton & Rovinelli, 1973) in order to determine the viability of the three methods to identify the 16 non-uniform DIF items described above. Non-uniform DIF was simulated by choosing different discrimination parameters for the two groups while keeping the difficulty parameters the same for the two groups. The DIF statistics values for the MH and LR procedures were obtained by using the program DICHODIF written by H. Jane Rogers and H.Swaminathan. The SIB statistics values were obtained using the program CSIBTEST (Li & Stout, 1993). The item parameter values for the 16 non-uniform DIF items are presented in Table 5.1.

In summary, DIF analyses were carried out with data sets simulated for four combinations of sample size, two levels of ability distribution differences, three levels of proportion of items containing DIF, four levels of DIF effect size, and four types of item. In all 384 conditions were studied to investigate non-uniform DIF. The data were replicated 100 times for each condition. The

power and Type I error rates of the three statistics were evaluated at the .05 and .01 levels of statistical significance.

In computing the MH DIF statistics, a two-stage procedure recommended by Holland and Thayer (1988) was adopted. With this procedure, items showing DIF using the total score as the matching criterion to group the examinees in the first-stage were excluded from forming the score groups in the second-stage. The two-stage procedure described above was not adopted while computing the SIB and LR DIF statistics.

Results

The Power Study

The results of the DIF analyses for the MH, SIB and LR procedures revealed in Tables 5.2 through 5.4 are summarized in the following sections.

To determine the effects of the five factors on the performance of the MH, SIB and LR procedures, an analysis of variance (ANOVA) was carried out. The dependent variable was the mean detection rates for the three procedures. The independent variables were the five different factors manipulated in the study. A review of ANOVA results presented in Table 5.2 shows that for all three procedures, three factors out of five, sample size, type of item and DIF effect size appear to have significant main effects that were common at .05 level of significance.

In addition, several two-way interaction effects observed were common for the three procedures. These were sample size by ability distribution, type of item by ability distribution, type of item by

DIF effect size, DIF effect size by ability distribution, DIF effect size by percent of DIF. For the MH procedure, there were interaction effects of type of item by sample size and type of item by percent of DIF. The significant interaction effects require some caution in the interpretations based on the main effects.

While the ANOVA table presented in Table 5.2 show a general trend in the results, a more detailed comparison of the three procedures is necessary to evaluate their performance. Table 5.3 and 5.4 present the effects of each of the five factors on the detection and Type I error rates of the three procedures averaged over all the conditions. The main findings of Tables 5.3 and 5.4 evaluated at .05 level of significance are as follows.

Effect of Sample Size. From Table 5.3 it is clear that the detection rates for the three procedures showed a steady increase for increase in sample size. In particular, the detection rates for the three procedures seemed to increase more for increase in the focal group sample size than for increase in the reference group sample size. The SIB procedure showed an increase of about 5% in detection rates over the LR procedure for all sample sizes. The detection rates for the MH procedure varied from about 31% to about 49% for the four sample sizes.

The Type I error rates presented in Table 5.4 show that they are within acceptable limits for the MH procedure for all sample sizes. They were higher than expected for the SIB and LR procedures with SIB showing an increase of about 0.5% over LR. For all three procedures, the Type I error rates were a slightly less for the lowest sample size than for other sample sizes.

Effect of Ability Distribution Difference. For all three procedures, the detection rates were higher when examinees were sampled from equal ability distribution than from unequal ability distribution. While the differences in detection rates for the two types of distribution were only about 2% to 3% for the MH and SIB procedures, it was much higher and was about 14% for the LR procedure.

For all three procedures, the Type I error rates were higher for unequal ability distribution than for equal ability distribution. While the Type 1 error rates were within acceptable limits for the MH procedure, they were higher than expected with SIB showing an increase of about 1% over LR.

Effect of Percent of Items Containing DIF. The detection rates for the MH and SIB procedures did not differ much whether the tests showed 10% or 20% of the items as DIF. For the LR procedure, there was an increase of about 4% for tests showing 10% of the items as DIF over tests showing 20% of the items as DIF.

The Type I error rates were seen to be within nominal limits for the MH procedure whether tests contained 0%, 10% or 20% items as DIF. They were slightly higher SIB and LR procedures, ranging up to about 8.6%.

Effect of Type of Item. The results show that overall, the detection rates for the four types of items were lowest for the MH procedure whereas, the correspondence between the detection rates for the SIB and LR procedures was very high. The detection rates for the SIB and LR procedures were highest for high discrimination/low difficulty items (about 88% and 90%) followed by high discrimination/medium difficulty items (about 77% and 70%). For the

MH procedure, these numbers were 66% and 22%. The detection rates for the SIB and LR procedures were lowest for medium difficulty/low discrimination items (about 47% and 44%) followed by high difficulty/low discrimination items (about 59% and 48%). For the MH procedure these numbers were about 15% and about 53%. The Type I error rates for the MH procedure were well within expected limits and higher then expected for the SIB and LR procedures (up to about 10.5% and 9.9%). For all three procedures, the Type I error rates were higher for highly discriminating items.

Effect of DIF Effect Size. The detection rates for the three procedures steadily increased for increase in the area values from .4 to 1.0. The lowest detection rates were observed for the MH procedure ranging from about 23% to about 50% for an increase in the area value of .4 to 1.0. For the SIB procedure, they ranged from 44% to 83% and for LR procedure, they ranged from 38% to 80%.

The ANOVA table (Table 5.2) show the results of ten two-way interaction effects among the five factors. One needs to be careful in interpreting the main effects of the ANOVA in view of the significant two-way interactions among the factors. To provide a more detailed comparison of the three methods due to interaction, a breakdown of detection rates by four such interactions are presented in Tables 5.5 through 5.8. The main findings of Tables 5.5 through 5.8 evaluated at .05 level of significance are summarized and presented below.

Effect of Sample Size by Ability Distribution. From Table 5.5 it is clear that the detection rates for all three procedures were

higher for equal ability distribution. The lowest detection rates were obtained for the MH procedure under both conditions.

Effect of Sample Size by Percent of DIF. Table 5.6 shows the results of the interaction effects of sample size by proportion of items containing DIF. Again, it is clear from this table that the detection rates were about the same for the MH procedure, increased marginally for the SIB procedure (about 2%) and decreased for the LR procedure (5% to 9%).

Effect of Type of Item by Ability Distribution. The interaction effects of the four different types of item by ability distribution are given in Table 5.7. The SIB and LR procedures showed a significant decrease in the detection rates for all types of item on the unequal ability distribution comparison. While the decrease was less for certain type of items they are much higher for other types of item. The pattern of detection was somewhat different for the MH procedure. For low and high difficulty items, the detection rates for the MH decreased when the ability distribution was unequal, whereas, for medium difficulty items, there was an increase in the detection rates for unequal ability distributions.

Effect of Type of Item by Percent of DIF. The results for the interaction effects of the type of item by proportion of items containing DIF (Table 5.8) show that the detection rates for the SIB procedure did not differ much for all types of items as the proportion of items containing DIF increased. For the LR procedure, the detection rates decreased for all types of items as the proportion of items showing DIF decreased. For the MH procedure, the detection

rates decreased for two types of items and increased for two types of items.

The results presented above indicate that while the SIB and LR procedures in general, are able to identify a high percentage of non-uniform DIF items, their inflated Type I error rates call for an adjustment to the values at the desired significance levels. To investigate such an adjustment, the Type I error rates of the SIB and LR procedures were evaluated at nine significance levels, viz., .05, .04, .03, .02, .01, .0075, .005, .0025 and .001 to determine the exact level of adjustment to the values at the desired level.

Tables 5.9 and 5.10 present the Type I error rates of the SIB and LR statistics at the nine significance levels. From these tables it is seen that for both procedures, the Type I error rates vary across all the three factors, viz., sample size, ability distribution and percent of DIF items. Figures 1 demonstrates graphically the results presented in Tables 5.9 and 5.10 for the sample size (500,200) and Figure 2 displays the results for the two ability distributions. It is clear from Figure 1 that the level of adjustment for α = .05 is to set it to α = .03 for both procedures. From Figure 2, it is seen that the impact of the equal and unequal ability distributions on the Type I error rates of the two procedures is different. While the Type I error rates are only slightly inflated for the two procedures for equal ability distributions, they are much higher for unequal ability distributions. From Figure 2 it is clear that for equal ability distribution, the level of adjustment for α = .05 is to set it at α = .034 for the SIB procedure and at α = .04 for the LR procedure. For unequal ability distribution, the level of adjustment for α = .05 is

to set it at α = .022 for the SIB procedure and at α = .025 for the LR procedure. The results presented above appear to be a possible solution to ensure that the Type I error rates are under control.

Discussion

Although uniform DIF occurs more often than non-uniform DIF in standardized tests, the investigation of non-uniform DIF under a variety of conditions was the main purpose of this study. The main findings of the study suggest that overall, there is high agreement between the SIB and LR procedures in detecting non-uniform DIF under most conditions. It is not surprising that the MH procedure was not capable of detecting non-uniform DIF under certain conditions because this procedure has been designed to detect uniform DIF only. As can be expected, all three procedures are affected by sample size. The detection rates for all three procedures increased when sample size increased. This is not surprising because, as sample size increases, the power of the DIF detection procedures also increases. Therefore, when differences are present, they are more likely to be detected.

The results of this appear to indicate that the detection rates depended more on the focal group sample size than the reference group sample size. Since this study investigated only four combinations of sample size, more research is needed in this area taking into consideration the ratio of the reference to the focal group sample size. The results of this study indicate that the power of the SIB and LR procedures in detecting non-uniform DIF were seen to be as high as 75% on average, for a focal group sample size of about 500.

The results also suggest that DIF effect size can have a significant effect on DIF detection procedures irrespective of the size and ratio of reference and focal group members. For all three procedures, the detection rates steadily increased when DIF effect sizes specified in terms of the areas between the ICCs for the two groups increased from 0.4 to 1.0. The lowest detection rates were seen to occur for the MH procedure varying between 23% and 50%. It is likely that items, showing small amounts of DIF may not be identified. However, in such cases test practitioners may not be concerned because the impact on test scores would be expected to be small.

The results support the findings of Rogers and Swaminathan (1993) that the type of item that the test is composed of is a significant factor influencing the detection rates of the DIF detection procedures. Their study comparing the MH and LR procedures showed that the MH procedure was not capable of detecting non-uniform DIF when the interaction was disordinal, i.e., when the ICCS of the two groups crossed in the middle of the ability distribution. Disordinal interactions occur with items of average difficulty. The MH statistic being a signed statistic, is sensitive to the direction of DIF. When the direction of DIF changes in the middle of the ability score distribution, negative differences in one part of the score distribution will cancel out against the positive differences in the other. Therefore, non-uniform DIF items of this form may not be detected by the MH procedure. The main purpose of including two items of average difficulty and one item each of low and high difficulty in this study was to investigate the performance of the three procedures to detect disordinal and ordinal interactions. The CRO-SIB procedure

developed recently for detecting non-uniform DIF has so far not been adequately tested for its capability to detect non-uniform DIF. Investigations of the new SIB procedure showed that it was equally powerful in detecting ordinal and disordinal interactions as the LR procedure under most of the studied conditions. For the two types of item included in this study where the interactions were ordinal (when the ICCS for the two groups crossed at the lower or upper end of the ability distribution) the performance of the MH procedure was comparable with the other two procedures.

In general, the detection rates for the SIB and LR procedures were highest for highly discriminating items with low difficulty followed by medium difficulty items. Low discrimination items with medium difficulty were least detected. For the MH procedure, the most significant factor to determine its capability to detect non-uniform DIF appears to be the type of item. While its performance appears to be comparable with the other two procedures in detecting DIF in easy and difficult items, it has limited use in the detection of DIF in average difficulty items. It appears that DIF in such items can be adequately detected by the SIB and the LR procedures.

The percentage of items showing DIF did not have a great impact on the DIF detection rates of the MH and SIB procedures, whereas, it did minimally affect the detection rates of the LR procedure (about 4%). This may be due to the two-stage procedure adopted in computing the MH statistic. Items identified as DIF in the first computations were removed when forming the score groups for computing the DIF statistics for the second time. The two-stage procedure was not adopted for the SIB and LR statistic for refinement and it is likely

that the results would have improved for both procedures if this had been incorporated.

The Type I error rates were within limits for the MH procedure. They were higher than expected for the SIB and LR procedures, with SIB results showing an overall increase of about 1% over LR results. In general, there appeared to be an increase in Type I error rates for the three procedures when the ability distribution differences increased or proportion of items containing DIF increased.

From the results presented above, it is clear that the SIB and LR procedures are equally effective in detecting non-uniform DIF in test items. However, the Type I error rates for both procedures are higher than the nominal level and call for an adjustment. This study indicated that the levels of adjustment vary with different conditions. The exact level of adjustment can be determined by evaluating the Type I error rates at a number of significance levels. The desired significance level can then be set to the adjusted significance level for the condition investigated. The MH procedure appears to have limited use in the detection of non-uniform DIF items which cross in the middle of the ability range. But this should not restrict its use in practice since it is seen to be very effective in detecting uniform DIF and some types of non-uniform DIF items.

In general, with an adjustment in the α level, either the CRO-SIB procedure or the logistic regression procedure can be used routinely for DIF detection. The CRO-SIB procedure is non-iterative and simple to implement. On the other hand, the logistic regression procedure is a general procedure and can be implemented readily using computer packages such as SPSS, BMDP, and SAS.

Table 5.1

Item Parameters Used to Generate Non-Uniform DIF Items

Item No.	Item Type	DIF Effect Size	Ref. b1	Foc. b2	Ref. a1	Foc. a2
1	Low b High a	.4	-1.50	-1.50	0.90	2.01
2		.6	-1.50	-1.50	0.70	1.97
3		.8	-1.50	-1.50	0.56	1.79
4		1.0	-1.50	-1.50	0.47	1.68
5	Medium b Low a	.4	0.00	0.00	0.50	0.72
6		.6	0.00	0.00	0.46	0.80
7		.8	0.00	0.00	0.43	0.91
8		1.0	0.00	0.00	0.40	1.03
9	Medium b High a	.4	0.00	0.00	0.90	2.01
10		.6	0.00	0.00	0.70	1.97
11		.8	0.00	0.00	0.56	1.79
12		1.0	0.00	0.00	0.47	1.68
13	High b Low a	.4	1.50	1.50	0.50	0.72
14		.6	1.50	1.50	0.46	0.80
15		.8	1.50	1.50	0.43	0.91
16		1.0	1.50	1.50	0.40	1.03

Table 5.2

Analysis of Variance of the Effects of all Factors on the Performance of the Mantel-Haenszel, Simultaneous Item Bias and the Logistic Regression Procedures

	MH		SIB		LR	
Factor	F	p	F	p	F	p
Main Effects						
Sample Size	29.46	.00*	60.43	.00*	46.34	.00*
Ability Distribution	1.41	.24	4.43	.01*	74.44	.00*
Percent DIF	0.00	.97	0.02	.89	6.14	.01*
Type of Item	281.98	.00*	215.94	.00*	187.27	.00*
DIF Effect Size	68.75	.00*	193.23	.00*	140.56	.00*
Interaction Effects						
Sample Size X Ability Distribution	4.78	.00*	0.69	.00*	0.89	.00*
Sample Size X Percent of DIF	0.02	.99	.44	.73	0.06	.98
Sample Size X Type of Item	4.30	.00*	1.66	.11	0.97	.47
Sample Size X DIF Effect Size	0.40	.93	0.87	.56	0.99	.45
Percent of DIF X Ability Distribution	0.46	.50	1.40	.24	3.20	.08
Type of Item X Ability Distribution	164.95	.00*	9.61	.00*	9.28	.00*
DIF Effect Size X Ability Distribution	4.64	.00*	4.80	00*	2.53	.05*
Percent of DIF X Type of Item	3.34	.02*	0.87	.46	0.72	.54
Percent of DIF X DIF Effect Size	4.01	.00*	9.10	.00*	5.20	.00*
Type of Item X DIF Effect Size	12.28	.00*	9.92	.00*	4.79	.00*

Table 5.3

Mean Percent Detection Rates of the Mantel-Haenszel, Simultaneous Item Bias and Logistic Regression Procedures Under all Conditions

Factor		MH		SIB		LR	
		α=.05 (%)	α=.01 (%)	α=.05 (%)	α=.01 (%)	α=.05 (%)	α=.01 (%)
Sample Size							
Ref	Foc						
500	200	31	18	58	41	52	33
500	500	41	29	72	57	69	55
1000	200	35	22	62	47	57	37
1000	500	49	37	79	68	75	63
Ability Distribution							
Equal		40	31	69	56	70	54
Unequal		38	22	66	51	56	40
Percent of DIF							
10%		39	26	68	53	65	49
20%		39	27	68	53	61	45
Type of Item							
Low b	High a	66	54	88	78	90	77
Medium b	Low a	15	6	47	30	44	26
Medium b	High a	22	12	77	63	70	53
High b	Low a	53	34	59	42	48	32
DIF Effect Size							
Area							
.4		23	12	44	28	38	21
.6		36	23	65	49	59	41
.8		47	33	78	64	74	58
1.0		50	38	83	71	80	67

Table 5.4

Mean Percent Type I Error Rates of the Mantel-Haenszel, Simultaneous Item Bias and Logistic Regression Procedures Under all Conditions

Factor		MH		SIB		LR	
		α=.05 (%)	α=.01 (%)	α=.05 (%)	α=.01 (%)	α=.05 (%)	α=.01 (%)
Sample Size							
Ref	Foc						
500	200	4.4	0.8	7.8	2.2	7.5	1.8
500	500	4.7	1.1	8.7	2.6	8.4	2.3
1000	200	4.5	1.0	8.2	2.5	8.5	2.2
1000	500	5.6	1.4	9.1	2.7	8.9	2.3
Ability Distribution							
Equal		4.1	0.9	7.0	1.8	6.1	1.4
Unequal		5.5	1.3	10.0	3.2	9.8	3.0
Percent of DIF							
0%		4.9	1.0	8.1	2.2	7.5	2.0
10%		4.8	1.1	8.4	2.5	8.6	2.3
20%		4.8	1.0	8.6	2.6	8.1	2.2
Type of Item							
Low b	High a	5.2	1.0	10.5	2.5	9.9	2.4
Medium b	Low a	4.4	1.1	7.6	1.8	7.5	1.3
Medium b	High a	5.3	1.2	9.4	1.9	8.8	2.2
High b	Low a	4.3	1.3	6.5	1.5	6.4	1.9

Table 5.5

Mean Percent Detection Rates of the Mantel-Haenszel, Simultaneous Item Bias and Logistic Regression Procedures for Sample Size by Ability Distribution

		Equal Ability Distribution						Unequal Ability Distribution					
		MH		SIB		LR		MH		SIB		LR	
	α =	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01
Factor		(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
Sample Size													
Ref	Foc												
500	200	34	24	59	43	58	38	27	13	57	38	46	27
500	500	43	35	74	61	77	66	40	24	69	54	60	44
1000	200	37	27	64	49	63	42	34	16	60	44	50	31
1000	500	45	39	79	69	81	70	53	34	79	68	69	56

Table 5.6

Mean Percent Detection Rates of the Mantel-Haenszel, Simultaneous Item Bias and Logistic Regression Procedures for Sample Size by Percent of DIF

		Percent of DIF Items = 10%						Percent of DIF Items = 20%					
		MH		SIB		LR		MH		SIB		LR	
	α =	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01
Factor		(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
Sample Size													
Ref	Foc												
500	200	31	18	58	41	54	34	31	19	57	40	49	31
500	500	41	29	71	57	71	58	41	30	72	58	67	52
1000	200	36	22	63	47	58	38	35	22	61	47	55	35
1000	500	49	37	78	68	78	67	49	37	80	69	73	60

Table 5.7

Mean Percent Detection Rates of the Mantel-Haenszel, Simultaneous Item Bias and Logistic Regression Procedures for Type of Item by Ability Distribution

		Equal Ability Distribution						Unequal Ability Distribution					
		MH		SIB		LR		MH		SIB		LR	
	α =	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01
Factor		(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
Type of Item													
Low b	High a	91	82	93	86	93	81	42	26	82	70	87	73
Med. b	Low a	6	3	44	27	48	30	23	10	50	32	39	22
Med. b	High a	3	1	81	68	84	68	41	23	73	58	57	37
High b	Low a	59	40	59	41	54	37	47	29	59	43	42	26

Table 5.8

Mean Percent Detection Rates of the Mantel-Haenszel, Simultaneous Item Bias and Logistic Regression Procedures for Type of Item by Percent of DIF

		Percent of DIF Items = 10%						Percent of DIF Items = 20%					
		MH		SIB		LR		MH		SIB		LR	
	α =	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01
Factor		(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
Type of Item													
Low b	High a	64	51	89	76	91	78	69	57	88	79	89	76
Med. b	Low a	13	5	46	28	44	26	17	8	48	31	43	26
Med. b	High a	26	14	77	64	73	55	19	10	77	63	68	51
High b	Low a	54	35	61	43	52	35	52	34	57	40	45	28

Table 5.9

Mean Percent Type I Error Rates of the Simultaneous Item Bias Statistic Computed at Nine Significance Levels

	α =	.05	.04	.03	.02	.01	.0075	.005	.0025	.001
Sample Size										
Ref	Foc									
500	200	7.8	6.4	5.1	3.6	2.2	1.8	1.3	0.8	0.4
500	500	8.7	7.1	5.8	4.3	2.6	2.0	1.3	0.9	0.5
1000	200	8.2	6.6	5.3	3.8	2.5	1.9	1.5	1.0	0.5
1000	500	9.1	7.2	5.9	4.1	2.7	2.0	1.5	1.0	0.5
Ability Distribution										
Equal		7.0	5.8	4.5	3.1	1.8	1.4	1.0	0.6	0.3
Unequal		10.0	7.8	6.4	4.8	3.2	2.5	1.9	1.2	0.6
Percent of DIF										
0%		8.1	6.5	5.1	3.7	2.2	1.6	1.1	0.6	0.3
10%		8.4	6.9	5.6	4.0	2.5	2.0	1.4	0.9	0.5
20%		8.5	6.7	5.3	3.8	2.6	1.9	1.4	0.8	0.4

Table 5.10

Mean Percent Type I Error Rates of the Logistic Regression Statistic Computed at Nine Significance Levels

	α =	.05	.04	.03	.02	.01	.0075	.005	.0025	.001
Sample Size										
Ref	Foc									
500	200	7.5	6.1	5.0	3.3	1.8	1.6	1.1	0.7	0.3
500	500	8.4	6.9	5.2	4.1	2.2	1.7	1.0	0.8	0.4
1000	200	8.5	7.0	5.3	4.0	2.4	1.7	1.1	0.9	0.4
1000	500	8.9	6.9	5.1	3.8	2.3	1.8	1.2	0.9	0.4
Ability Distribution										
Equal		6.1	5.1	4.0	2.8	1.4	1.1	0.8	0.4	0.2
Unequal		9.8	7.4	6.1	4.2	3.0	2.3	1.6	1.0	0.4
Percent of DIF										
0%		7.5	5.8	4.8	3.3	2.0	1.3	0.9	0.4	0.2
10%		8.6	6.0	5.1	3.6	2.3	1.5	1.0	0.6	0.3
20%		8.1	5.7	5.0	3.5	2.2	1.5	1.0	0.5	0.3

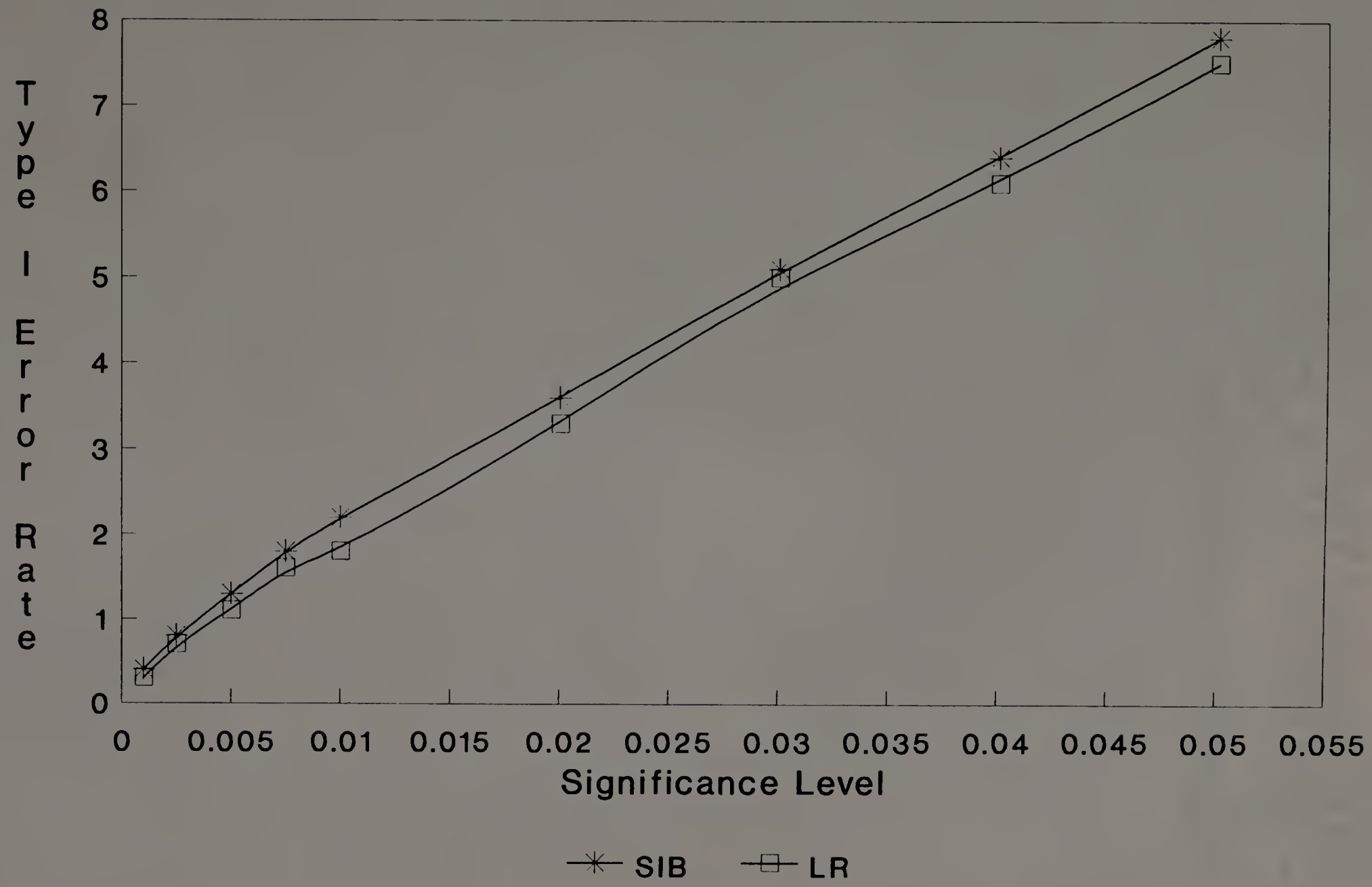

Figure 1 Type I Error Rates of the SIB and LR Procedures for Sample Size: Reference Group = 500; Focal Group = 200.

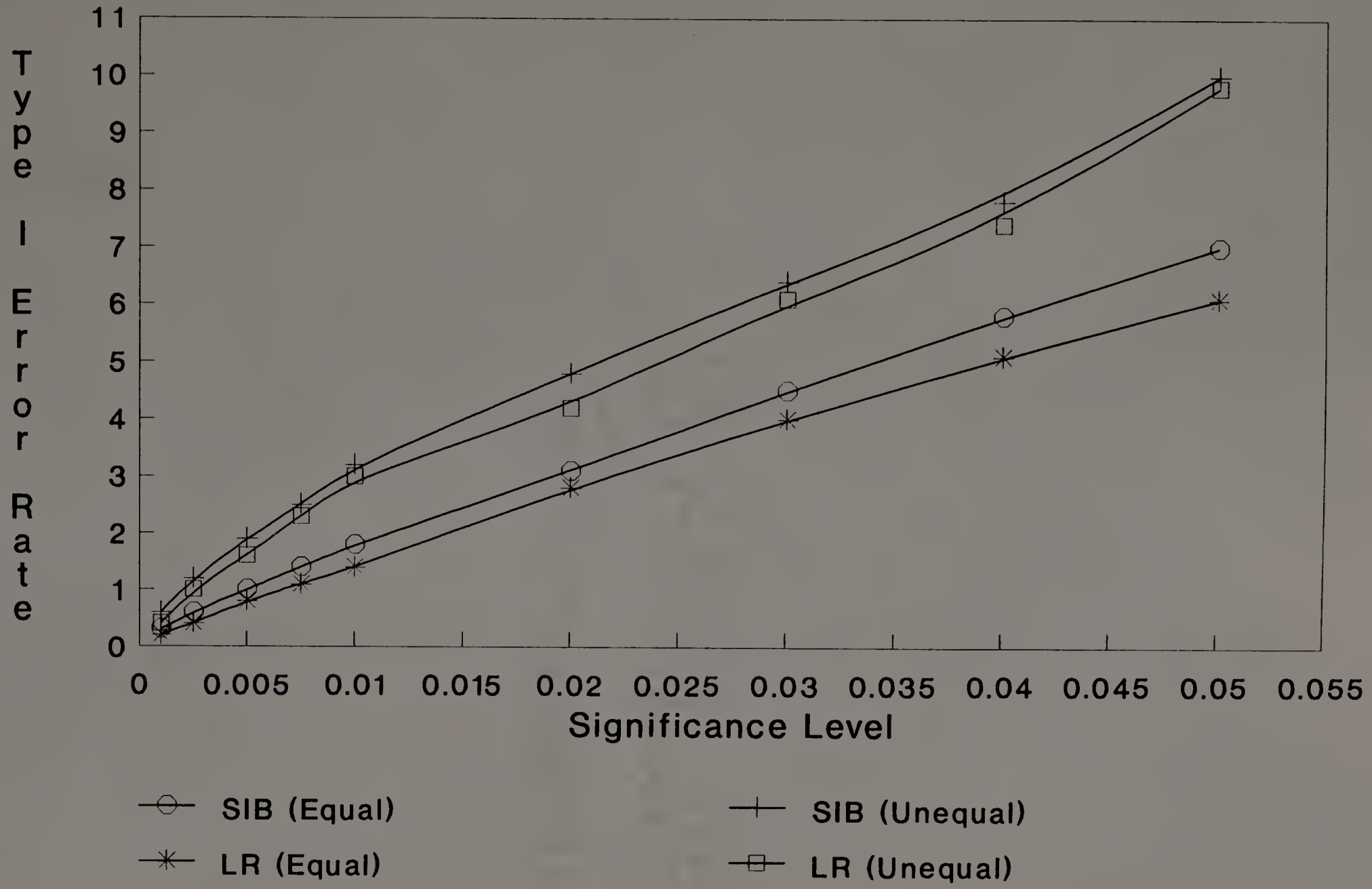

Figure 2 Type I Error Rates of the SIB and LR Procedures for Equal and Unequal Ability Distributions.

C H A P T E R VI

CONCLUSIONS

Summary

The performance of three currently popular procedures DIF detection procedures, the Mantel-Haenszel (MH), the Simultaneous Item Bias (SIB) and the Logistic Regression (LR) procedures were examined in three different studies.

The first study compared two non-parametric procedures for detecting DIF, the MH and the SIB procedures, with respect to their Type I error rates and power to detect uniform DIF. Data for the study were simulated to reflect a variety of conditions. In all, 1296 conditions were studied with data sets simulated for nine combinations of sample size, three level of ability distribution differences between the focal and the reference groups, two levels of the proportion of DIF items in the test, four levels of DIF effect sizes, and six types of item.

Investigations of the power of the SIB and the MH procedures revealed that both procedures were almost equally powerful in detecting uniform DIF under most of the studied conditions. As expected, the statistical power of both procedures increased for increase in sample size and DIF effect size. Both procedures appeared to have sufficient power (about 85%) to detect DIF for a focal group sample size of about 300 examinees irrespective of the reference group sample size. These results are encouraging because in practice, it is not always possible to have large samples of focal group examinees.

One significant finding was the impact of the ability distribution differences on the detection rates of the two procedures. While both procedures were almost equally effective in detecting DIF when examinees were sampled from equal ability distributions, the performance of the SIB was superior to that of MH for examinees sampled from unequal ability distributions under certain conditions. For certain types of items, the SIB procedure was able to detect about 25% more items as DIF than the MH procedure when there were differences in the ability distributions. In practical settings, these results suggest that the SIB procedure could be more useful than other DIF detection procedures for comparisons between groups of differing abilities. Further research would be needed to confirm these results.

The proportion of items showing DIF did not have an impact on the detection rates of both procedures possibly due to the two-stage procedure adopted in computing both the DIF statistics. The type of item appeared to have a significant impact on the detection rates of the two procedures. The results showed that the detection rates of the two procedures decreased as the difficulty levels of the items increased and increased as the discrimination levels of the items increased.

The Type I error rates of the MH statistic were within the expected limits under almost all conditions, whereas, they were slightly inflated for the SIB procedure for many conditions. In conclusion, the results of this research show that both the MH and the SIB procedures are viable procedures for detecting uniform DIF. In practice, the SIB procedure can be used when there are ability

distribution differences between the two groups. If the practitioner is concerned about the Type I error rates, then the MH procedure is recommended for use in DIF studies.

This main purpose of the second study was the investigation of the distributional properties of the MH and the SIB procedures. Previous studies on the distributional properties of the (MH) statistic have revealed that the statistic as recommended by Holland and Thayer (1988) does not have the approximate chi-square distribution. To study the impact of the continuity correction in the computation of the MH statistic, investigations of the distributional properties of the MH statistic included two variations of the statistic (with and without the continuity correction). Although the power and Type I error rates of the MH and the SIB statistic have already been examined in the first study, the power analyses were repeated in this study so that the performance of the MH statistic without the continuity correction could be compared with the SIB and the original MH statistic.

The distribution study was conducted with simulated data sets for 81 conditions obtained by crossing nine combinations of sample size with six types of items. In the power study, the conditions specified in the first study were examined again with the results of the MH statistic without the continuity correction included in the analyses.

The results of the distribution study suggested that for almost all the studied conditions, the empirical distributions of the SIB statistic approached their expected distributions for a sample size of 200 or more in both the groups. Both MH statistics did not have their

expected distributions for any of 81 studied conditions. However, the results on the investigations of the distribution of MH statistic without the continuity correction were more acceptable than those from the MH statistic with the continuity correction for all the studied conditions.

The results from the power study were quite similar to those reported in the first study. In addition, the results from the MH statistic without the continuity correction were close to the SIB results which had higher identification rates under many conditions. Investigations of the Type I error rates of the three statistics showed that they within the nominal limits and conservative for the MH statistic with the continuity correction, within limits under most conditions for the MH statistic without the continuity correction and slightly inflated for the SIB statistic.

In conclusion, this study recommends the use of the SIB statistic for detecting uniform DIF due to its adherence to its distributional assumptions if the Type I error rates are tolerable in practice. Test practitioners should use of the MH statistic without the continuity correction in place of the MH statistic with the continuity correction is recommended if the Type I error rates need to be in control in practical settings.

The purpose of the third study was to compare three procedures, the MH, the SIB and the LR procedures, with respect to their Type I error rates and power to detect non-uniform DIF. Previous research on the detection of non-uniform DIF have shown that the LR procedure is superior to all other procedures in detecting non-uniform DIF. The SIB procedure has been investigated in several studies previously for

its capability to detect uniform DIF. A new SIB procedure to detect non-uniform DIF was developed recently and has not been extensively studied so far. Because the LR procedure is the only known procedure at present which is very effective in detecting non-uniform DIF, a comparative study of the new SIB procedure with the MH and the LR procedures was the focus of this third research.

Data sets were simulated for 384 conditions by manipulating four combinations of sample size, two levels of ability distribution differences between the focal and the reference groups, three levels of the proportion of DIF items in the test, four levels of DIF effect sizes, and four types of item.

The main findings of the study revealed that both SIB and LR procedures were almost equally powerful in detecting non-uniform DIF under most conditions. The statistical power of both procedures was about 70% for a focal group sample size of about 500 examinees. The MH procedure was not very effective in identifying non-uniform DIF when the interaction between the ability variable and group membership was disordinal (for example, items of medium difficulty).

Investigations of the Type I error rates of the three statistics showed that they were within the nominal limits for the MH procedure and higher than the nominal levels for the SIB and LR procedures. The SIB results showed an overall increase of about 1% over the LR results. A practical solution to control the Type I error rates of the SIB and the LR statistics has been offered in this study. It calls for an adjustment in the significance levels by evaluating the Type I error rates at a number of significance levels. The desired

significance level could be then set to the adjusted level for the condition investigated.

In conclusion, the results of this study show that both the SIB and the LR procedures are viable methods for detecting non-uniform DIF although the Type I error rates of both procedures need an adjustment recommended in the study. The MH procedure could also be useful in detecting non-uniform DIF only under certain conditions.

Implications for Practice

The results from the three studies described above have several implications in practice. First, when measures of statistical significance of the DIF detection procedures are used for screening items for DIF, they will be effective only when they satisfy the distributional assumptions on which they are based. In practice, the use of the MH statistic without the continuity correction is strongly recommended if DIF analyses are carried out with the MH procedure.

Practitioners should be aware that the power of all DIF detection procedures increase when sample sizes increase. Therefore, it would be more preferable to investigate DIF as far as possible with large sample sizes. But when in practice samples sizes are limited, a sample size of about 300 in each group may be sufficient to provide power (about 85%) to detect items showing uniform DIF and about 500 in each group would be required to provide power (about 72%) to detect non-uniform DIF. Focal group samples less than those mentioned above may be insufficient for DIF detection purposes.

This study confirmed the findings of existing research that the size of DIF can affect DIF detection rates. When the size of DIF in

terms of the area between the ICCs for the two groups was only .4, the power of the DIF detection procedures were seen to be very low for detecting both types of DIF, but increased markedly as the size of DIF increased to the area value 1.0. Practitioners should be reminded that items which show small amounts of DIF can go undetected especially when sample sizes are small. However, it can be argued that in such cases, the DIF may be so small that it would make very little practical difference on test score.

Previous research have demonstrated the influence of different types of item on DIF detection rates. The power of DIF detection procedures are seen to decrease when the difficulty levels of the items increase. They tend to increase when the discrimination levels of the items increase. Standardized achievement tests are likely to contain more moderate difficulty items than low or high difficulty items. Even if a few highly difficult items are present in the tests, they will affect only a small number of examinees likely to be present at the extreme ends of the ability continuum. Therefore, the presence of DIF not identified in these items may not be a matter of great concern.

Although uniform DIF is more common than non-uniform DIF, the existence of non-uniform DIF in test items should not be ignored. The SIB and LR have demonstrated their capability to detect non-uniform with a reasonable amount of accuracy under certain conditions and therefore, can be very useful in practice for such types of items. Although the MH procedure is seen to provide satisfactory results for tests that are very easy or very difficult in detecting non-uniform

DIF, it might fail to identify items of moderate difficulty even when there are large amounts of DIF.

In practice, a feasible alternative suggested by Hambleton and Rogers (1989) may be applied to identify items showing non-uniform DIF. The authors suggest a routine comparison of the direction of differences in the p-values for the two groups of interest across the score groups. If the direction of the difference favored one group at test scores below a certain test score and favored the other group above the test score, then non-uniform DIF is likely to be present.

The Type I error rates seem to be a problem with the SIB and LR statistics. A practical solution to control the Type I error rates of the SIB and LR statistics has been offered in the study. It calls for an adjustment in the Type I error rates at a number of significance levels. The desired significance level could be then set at the adjusted level for the condition investigated.

In practice, DIF comparisons are commonly carried out with examinees sampled from equal ability distributions or examinees sampled from unequal ability distributions. The SIB procedure seems to be more effective than other DIF detection procedures when there are ability distribution differences in DIF studies. The impact of ability distribution differences is minimal on the SIB procedure. This is due to the regression correction effected in the SIB statistics which adjusts the means of the studied subtest for ability distribution differences if they exist. This study recommends the use of the SIB statistic when there are ability distribution differences between the two groups. Again the Type I error rates seem to be a problem for the SIB statistic and needs the adjustment suggested earlier.

The two-stage procedure is recommended in practice since the percentage of items showing DIF did not affect the detection rates of the SIB and the MH procedures when this procedural variation was incorporated into the two statistics. The two-stage procedure is also very easy to implement in practice.

Directions for Future Research

Currently three of the most popular methods for detecting DIF are the MH, SIB and the LR procedures. While a large number of studies have examined the influence of several issues (for example, sample size and other factors) on the performance of the MH procedure, existing research on the two relatively new SIB and the LR procedures are much less. The three studies described above have examined a number of issues not previously studied to add more information to the existing research on these three procedures. Many areas of research were beyond the scope of this study and merit further investigation.

One area of future research on DIF would be to focus more attention on the issue of sample sizes. The information currently available to test administrators regarding the appropriate sample size in DIF studies is inadequate for making decisions about the statistical screening of item bias. This study examining the issue of sample sizes indicate that sample sizes less than 300 in each group may not be enough for the DIF detection techniques to perform with a reasonable amount of accuracy. Therefore, more research should focus on manipulating the ratio of sample sizes in the reference and the focal groups.

Another area of future research should focus on the ability distributions of the reference and the focal groups. In practice, DIF studies are often required to be undertaken with examinees from equal ability distributions when comparing gender groups or examinees from unequal ability distributions when comparing ethnic groups. These comparisons can be made more effectively in practice if reliable DIF statistics are available to determine the impact of the ability distribution factor on DIF detection procedures. Research should continue in this area by varying the means and the standard deviations of the two groups to represent the different distributions found in practical settings.

The present DIF studies usually use measures of either statistical significance or estimates of DIF effect size of the DIF statistics to identify differently functioning items. It would be more informative in practice to examine both the measures when DIF studies are conducted.

Very limited number of existing research have examined the distributional properties of the statistics used in DIF studies. For a statistic to be an effective indicator of the presence of DIF in test items, its distributional assumptions must hold. Previous research on the distributional assumptions of the MH statistic have not produced satisfactory results for many conditions. Investigations of a variation of this statistic obtained by removing the continuity correction appear to meet the distributional assumptions to a greater degree than without the continuity correction, at least for practical purposes. More research needs to be done especially with respect to

its Type I error rates before this procedural variation could be recommended for implementation in practice.

The distribution studies currently available on the DIF detection procedures have only examined the distributional properties of the test statistics with data generated for examinees of equal ability in both groups. Future research should also focus on examining the distributional assumptions of these statistics with data generated for examinees of unequal ability in both groups to determine the validity of the statistic for such comparisons in practice.

In most currently available DIF studies, the total score is used to match the focal and the reference group examinees to form score groups. The results from these studies are valid only when tests are unidimensional. When tests are intentionally multidimensional created to measure complex skills, DIF studies conducted with unidimensional assumptions might yield invalid results. Ackerman (1992) has shown that the identification of a "valid subtest" of items to be used as the matching criterion instead of the total score could reduce false positive error rates. Such studies can be undertaken with DIF detection procedures only when it is possible to identify "valid subtests" in the available data. Identification of such "valid subtests" may be difficult in most DIF studies. Future research should focus on the issue of multidimensionality in DIF studies using procedures that can accommodate the option of selecting a "valid subtest" of items. In this context, the use of SIB and the LR procedures will be very useful. Both procedures have the flexibility to condition on one or more test or subtest score.

Conclusions

With a proliferation of methods available for detecting DIF, empirical comparisons of these methods will be useful in generalizing the results to offer guidelines for test developers about the advantages and disadvantages of various DIF detection techniques. With this purpose in mind, this study compared the MH, SIB and the LR procedures.

In general, the results from a comparison of the three procedures for DIF detection purposes showed some of the merits and demerits of the three procedures. The MH procedure is very effective in detecting uniform DIF, but has limited use in the detection of non-uniform DIF items of moderate difficulty. Despite the violations of the distributional assumptions of this procedure under most conditions, its Type I error rates throughout the study were seen to be within the nominal levels.

The SIB procedure has demonstrated its capability to detect uniform DIF as effectively as the MH procedure and detect non-uniform DIF as effectively as the LR procedure. It is very robust with respect to its distributional assumptions, but the inflation of Type I error rates observed throughout this study can be a limitation to this procedure.

The LR procedure's investigations were limited to non-uniform DIF in this study. It performed as effectively as the SIB procedure while keeping the Type I error rates a little lower than the SIB procedure. Further research is recommended.

All three procedures are theoretically sound, computationally non-intensive, and effective with small sample sizes. While the MH

and the SIB procedures are non-iterative and can easily be implemented in practice, the LR is a more general procedure that can be implemented with computer packages such as SPSS, SAS and BMDP.

REFERENCES

Ackerman, T. A. (1992, April). An investigation of the relationship between reliability, power and Type 1 error rate of the Mantel-Haenszel and the simultaneous item bias procedures. Paper presented at the meeting of NCME, San Francisco, CA.

Angoff, W. H. (1982). The use of difficulty and discrimination indices in the identification of biased test items. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 96-116). Baltimore, MD: Johns Hopkins University Press

Angoff, W. H. & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10, 95-106.

Baker, F. B. (1981). A criticism of Scheuneman's item bias technique. Journal of Educational Measurement, 18, 56-62.

Berk, R. A. (Ed.). (1982). Handbook of methods for detecting test bias. Baltimore, MD: Johns Hopkins University Press.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Bock, R. D. (1975). Multivariate statistical methods. New York: McGraw-Hill.

Camilli, G. A. (1979). A critique of the chi-square method for assessing item bias. Unpublished paper, Laboratory of Educational Research, University of Colorado, Boulder.

Camilli, G. A., & Shepard, L. A. (1987). The inadequacy of ANOVA for detecting test bias. Journal of Educational Statistics, 12, 253-260.

Cardell, G. L. & Coffman, W. E. (1964). A method for comparing the performance of different groups on the items in a test (Research Bulletin RB - 64-61). Princeton, NJ: Educational Testing Service.

Cleary, T. A. & Hilton, T. L. (1968). An investigation of item bias. Educational and Psychological Measurement, 28, 61-75.

Coffman, W. E. (1961). Sex differences in responses to items in an aptitude test. In Eighteenth Yearbook of the National Council on Measurement in Education, 117-124.

Coffman, W. E. (1963). Evidence of cultural factors in responses of African students to items in an American test of scholastic aptitude (Research and Development Reports). New York: College Entrance Examination Board.

Crocker L., & Algina, J. (1986). Introduction to classical and modern test theory. Orlando, FL: Holt, Rinehart & Winston, Inc.

Dorans, N. J., Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude test. Journal of Educational Measurement, 23, 355-368.

Echternacht, G. (1974). A quick method for determining test bias. Educational and Psychological Measurement, 34, 271-280.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items; Comparison of IRT area and Mantel-Haenszel methods. Applied Measurement in Education, 2(4), 313-334.

Hambleton, R. K., & Rovinelli, R. (1973). A FORTRAN IV program for generating response data for logistic models. Behavioral Science, 18, 73-74

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Hingham, MA: Kluwer Academic Publishing.

Hills, J. R. (1989). Screening for potentially biased items in testing programs. Educational Measurement: Issues and Practice, 8(4), 5-11.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and Mantel-Haenszel procedure. In H. Wainer & H I. Braun (Eds.), Test Validity. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hoover, H. D., & Kolen, M. J. (1984). The reliability of six item bias indices. Applied Psychological Measurement, 8, 173-181.

Ironson, G. H. (1982). Use of the chi-square and latent trait approaches for detecting item bias. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 117-160). Baltimore, MD: Johns Hopkins University Press.

Ironson, G. H., Homan, S., Willis, R., & Signer, B. (1984). The validity of item bias techniques with math word problems. Applied Psychological Measurement, 8, 391-396.

Kingston, N., Leary, L., Wightman, L. (1988). An exploratory study of the applicability of item response methods to the Graduate Management Admission Test (GMAT Occasional Papers). Princeton, NJ: Graduate Management Admission Council.

Li, H., & Stout, W. F. (1993, April). A new procedure for detecting crossing DIF/bias. Paper presented at the meeting of AERA, Atlanta.

Linn R. L., & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. Journal of Educational Measurement, 18, 109-118.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). An investigation of item bias in a test of reading comprehension. Educational Measurement: Issues and Practice, 6(2), 13-17.

Lord, F. M. (1980). Applications of item response theory. Hillside, NJ: Lawrence Erlbaum.

Maclaughlin, M. E., & Drasgow, F. (1987). Lord's chi-square test of item bias with estimated and with known person parameters. Applied Psychological Measurement, 11, 161-173.

Marascuilo, L. A., & Slaughter, R. E. (1981). Statistical procedures for identifying possible sources of item bias based on chi-square statistics. Journal of Educational Measurement, 18, 229-248.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. Educational and Psychological Measurement, 52, 443-451.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. Journal of Educational Statistics, 7, 105-118.

Narayanan, P. & Swaminathan, H. (1993, April). Performance of the Mantel-Haenszel and Simultaneous Item Bias procedures for detecting differential item functioning. Paper presented at the meeting of NAME, Atlanta, GA.

Petersen, N. S., & Novick, M. R. (1976). An evaluation of some models for culture-fair selection. Journal of Educational Measurement, 13, 3-29.

Plake, B. S. (1980). A comparison of a statistical and subjective procedure to ascertain item validity: one step in the test validation process. Educational and Psychological Measurement, 40, 397-404.

Raju, N. S. (1988). The area between two item characteristic curves. Psychometrika, 53, 495-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. Applied Psychological Measurement, 14(2), 197-207.

Raju, N. S., Bode, R. K., & Larsen, V. S. (1989). An empirical assessment of the Mantel-Haenszel statistic for studying differential item performance. Applied Measurement in Education, 2, 1-13.

Rogers, H. L. (1989). A logistic regression procedure for detecting item bias. Unpublished Doctoral Dissertation, University of Massachusetts at Amherst, MA.

Rogers, H. L., & Hambleton, R. K. (1989). Evaluation of computer simulated baseline statistics for use in item bias studies. Educational and Psychological Measurement, 49, 355-369.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. Applied Psychological Measurement, 17, 105-116.

Roussos, L. A., & Stout, W. F. (1993, April) Simulation studies of effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. Paper presented at the meeting of AERA, Atlanta.

Rudner, L. M. (1977). Efforts toward the development of unbiased selection and assessment instruments. Paper presented at the Third International Symposium on Educational Testing, University of Leiden, The Netherlands.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980a). Biased item detection techniques. Journal of Educational Statistics, 5, 213-233.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980b). A Monte Carlo comparison of seven biased item detection techniques. Journal of Educational Measurement, 17, 1-10.

Scheuneman, J. D. (1979). A method of assessing bias in test items. Journal of Educational Measurement, 16, 143-152.

Scheuneman, J. D., & Bleistein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. Applied Measurement in Education, 2, 255-275.

Shealy, R. & Stout, W. F. (1991, April). An item response theory model for test bias. Paper presented at the meeting of American Educational Research Association, Chicago.

Shealy, R. & Stout, W. F. (1993). A model based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. Psychometrika, 58, 159-194.

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting item bias with both internal and external ability criteria. Journal of Educational Statistics, 6, 317-375.

Shepard, L., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. Journal of Educational Statistics, 9, 93-128.

Shepard, L., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22, 77-105.

Sinnott, L. T. (1980). Differences in item performance across groups (ETS research Report 80-19). Princeton, NJ: Educational Testing service.

Subkoviak, M. J., Mack, J. S., Ironson, G. H., & Craig, R. D. (1984). Empirical comparison of selected item bias detection procedures with bias manipulation. Journal of Educational Measurement, 21, 49-58.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27(4), 361-370.

Van der Flier, H., Mellenbergh, G. J., Ader, H. J., & Wihn, M. (1984). An iterative item bias detection method. Journal of Educational Measurement, 21, 131-145.

Wise, L. L. (1987). Differential item difficulty indicators in small samples. Paper presented at the meeting of AERA, Washington.

Wright, D. J. (1986, April). An empirical comparison of the Mantel-Haenszel and Standardization methods for detecting differential item performance. Paper presented the meeting of NCME, San Francisco, CA.

Wright, B. D., Mead, R., & Draba, R. (1976). Detection and correcting test item bias with a logistic response model (Research Memorandum No. 22). Chicago: University of Chicago, Department of Education, Statistical Laboratory.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide. Journal of Educational Measurement, 3, 185-197.